

Asymptotically Optimal Matching of Multiple Sequences to Source Distributions and Training Sequences

Jayakrishnan Unnikrishnan

Audiovisual Communications Laboratory, School of Computer and Communication Sciences

Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

Email: jay.unnikrishnan@epfl.ch

Abstract

Consider a finite set of sources, each producing i.i.d. observations that follow a unique probability distribution on a finite alphabet. We study the problem of matching a finite set of observed sequences to the set of sources under the constraint that the observed sequences are produced by distinct sources. In general, the number of sequences N may be different from the number of sources M , and only some $K \leq \min\{M, N\}$ of the observed sequences may be produced by a source from the set of sources of interest. We consider two versions of the problem – one in which the probability laws of the sources are known, and another in which the probability laws of the sources are unspecified but one training sequence from each of the sources is available. We show that both these problems can be solved using a sequence of tests that are allowed to produce “no-match” decisions. The tests ensure exponential decay of the probabilities of incorrect matching as the sequence lengths increase, and minimize the “no-match” decisions. Both tests can be implemented using variants of the minimum weight matching algorithm applied to a weighted bipartite graph. We also compare the performances obtained by using these tests with those obtained by using tests that do not take into account the constraint that the sequences are produced by distinct sources. For the version of the problem in which the probability laws of the sources are known, we compute the rejection exponents and error exponents of the tests and show that tests that make use of the constraint have better exponents than tests that do not make use of this information.

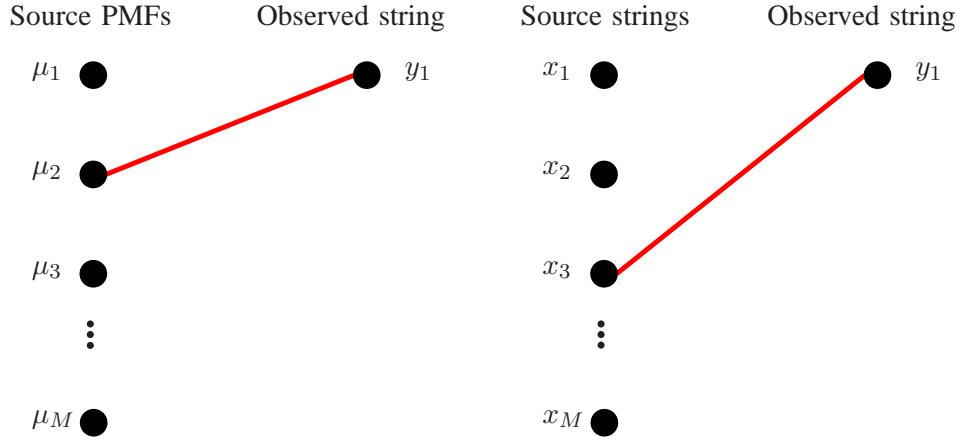
Portions of this paper were presented in part at the 51st Annual Allerton Conference on Communication, Control, and Computing, Monticello, Illinois, October 2013 [1].

I. INTRODUCTION

Classical multi-hypothesis testing [2] addresses the following problem: Given probability distributions of M sources and one observation sequence (or string), decide which of the M sources produced the sequence. Classical statistical classification [3] also addresses the same problem with the only difference that the probability distributions of the sources are not known exactly, but instead, have to be estimated from training sequences produced by the sources. Figure 1 illustrates these classical problems. In this paper we study a generalization of these problems which is relevant in applications like de-anonymization of anonymized data [1]. Instead of one observation sequence, suppose that you are given N observation sequences, subject to the constraint that each sequence is produced by a distinct source. We consider the task of matching the sequences to the correct sources that produced them, as illustrated in Figure 2. Focusing on finite alphabet sources, we study these matching problems as composite hypothesis testing problems. We refer to the first problem, in which the distributions of the sources are known, as the *matching problem with known sources*, and the second problem in which only training sequences under the sources are given, as the *matching problem with unknown sources*. We obtain solutions to both these problems that are asymptotically optimal in error probability as the length of the sequences increases to infinity.

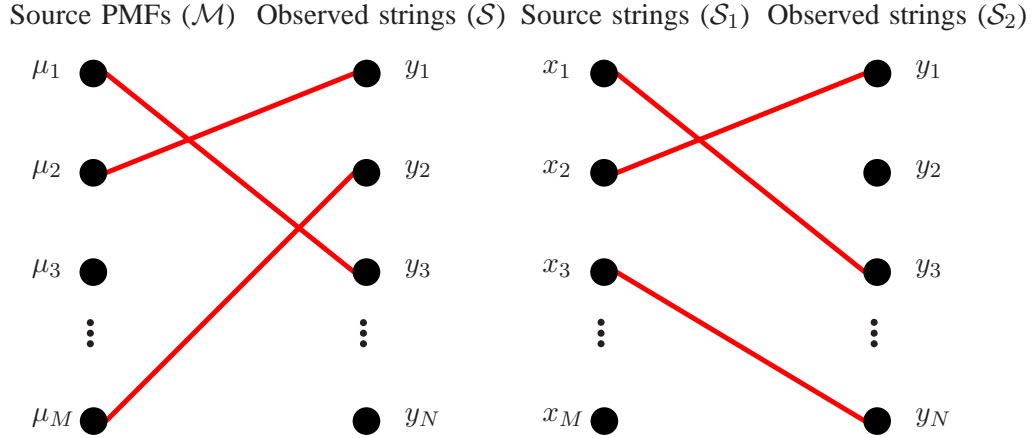
The main difference between these problems and the standard multi-hypothesis testing and classification problems is the constraint that the observation sequences are produced by distinct sources. It is clear that in the absence of such a constraint, these problems are just repeated versions of the standard problems. The constraint adds more structure to the solution and leads to an improvement in classification accuracy. We use large deviations analysis to quantify the improvement in performance in terms of the asymptotic rate of decay of the error probabilities and the probabilities of rejection of the optimal tests with and without the constraints. We obtain asymptotically optimal solutions to these matching problems using a generalization of the approach of Gutman [4], who solved the classical statistical classification problem.

Our primary motivation for studying these problems comes from studies on privacy of anonymized databases. In recent years, many datasets containing information about individuals have been released into public domain in order to provide open access to statistics or to facilitate data mining research. Often these databases are *anonymized* by suppressing identifiers that reveal the identities of the users, like names or social security numbers. Nevertheless, recent research (see, e.g., [5], [6]) has revealed that the privacy offered by such anonymized databases may be compromised if an adversary correlates the revealed information with publicly available databases. In our recent work [1], we studied the privacy



(a) Multihypothesis testing: Match observed string to correct source probability mass function (PMF) (b) Statistical classification: Match observed string to the training string from the correct source

Fig. 1. Illustration of the matching tasks in classical multihypothesis testing and statistical classification



(a) Matching problem (**P1**): Match observed strings to correct source PMFs (b) Matching problem (**P2**): Match observed strings to the training strings from the correct sources

Fig. 2. Illustration of the matching tasks in the problems studied in this paper

of anonymized user statistics in the presence of auxiliary information. We showed that anonymized statistical information about a set of users can be easily de-anonymized by an adversary who has access to independent auxiliary observations about the users. The task of the adversary is to match the auxiliary information to the anonymized statistics, which is exactly the problem that is studied in the current paper. Another related application of the matching problem is in matching statistical profiles of users obtained

from two different sources. For example, one could obtain location statistics of a set of users either from connections to WiFi access points, or from connections to mobile towers. It is interesting to try to match users across these two datasets, using the statistics of their location patterns. Alternatively, the user statistics could be the frequency distributions of words used by users on two different blog websites. The matching task then is to identify users who have accounts on both websites. Matching users across two different datasets increases the net information available about the users which in turn can be used to improve accuracy of targeted services.

Asymptotically optimal hypothesis testing has a long history in literature (see e.g., [7]–[10]). However, hypothesis testing of multiple sequences under the constraint that each sequence is produced by a distinct source, has been studied only rarely. The prior knowledge of the constraint on the sequences is expected to improve the accuracy of the hypothesis test. However, the task of identifying the optimal solution is now much more complicated as there are a combinatorial number of hypotheses. It is not immediately clear what is the best strategy to adopt. A naive strategy is to try to classify each sequence individually; but that is not expected to yield high accuracy as the constraints are not intelligently exploited. In [11, Ch. 10] the matching problem with known distributions was studied for the special case of $M = N$, where the analysis was performed by reducing the problem to a multi-hypothesis testing problem. The same problem was solved in [12] for $M = N = 2$ under a different optimality criterion from that used in this paper. In the first part of this paper we study this problem under a different optimality criterion, and identify an optimal solution for general M and N . We also provide a quantitative comparison of the performance obtained with our optimal solution, with that of a test that ignores the constraint on the sequences. These results quantify the improvement in performance that can be obtained by exploiting the constraint on the sequences. In the second part of this paper we study the problem of matching one set of sequences to another set of sequences. The approach we adopt in most of this paper is a generalization of that adopted by Gutman in [4], who solved the matching problem with unknown source distributions for $N = 1$. Gutman showed that if a “no-match” decision is allowed, it is possible to guarantee exponential decay of all misclassification probabilities at a desired rate. In the second part of this paper we simplify the structure of Gutman’s solution and show that his method can be generalized to solve the matching problem with unknown source distributions for general N . We also demonstrate that although there are a combinatorial number of hypotheses for these problems, simple polynomial-time algorithms can be used to identify the optimal solutions.

The rest of the paper is organized as follows. After introducing our notation, we state the problems in mathematical form in Section II. We present our solution to the generalization of the hypothesis testing

problem in Section III and our solution to the generalization of the statistical classification problem in Section IV. In addition to identifying the optimal solutions, we also compare the performances of these solutions with those of solutions that do not explicitly take into account the constraint on the distinctness of the sources that produced the sequences. We discuss practical aspects of implementing the test and conclude in Section V. For ease of reading, we relegate proofs of all results to the appendix.

Notation: For a finite alphabet Z , we use $\mathcal{P}(Z)$ to denote the set of all probability distributions defined on Z . We interchangeably use the words sequence and string to refer to an ordered list of elements from Z . For any string $s \in Z^n$, we use $\Gamma_s \in \mathcal{P}(Z)$ to denote the empirical distribution of the string defined as

$$\Gamma_s(z) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}\{s_i = z\}, z \in Z.$$

For $\mu \in \mathcal{P}(Z)$ and $z \in Z$ we use $\mu(z)$ to denote the probability mass at z under μ . For a string $s \in Z^n$, we use $\mu(s)$ to denote the probability of observing s at the output of a source that generates n observations i.i.d. according to law μ . We use $H(\mu)$ to denote the Shannon entropy

$$H(\mu) = \sum_{z \in Z} -\mu(z) \log \mu(z).$$

For $\nu, \mu \in \mathcal{P}(Z)$ we use

$$D(\nu \parallel \mu) = \sum_{z \in Z} \nu(z) \log \frac{\nu(z)}{\mu(z)},$$

to denote the Kullback-Leibler divergence between probability distributions ν and μ . Throughout the paper we use \log to refer to logarithm to the base 2. We use $[N] := \{1, 2, \dots, N\}$ and $\text{Sym}([N])$ to denote the set of all permutations on $[N]$, i.e., if $\sigma \in \text{Sym}([N])$ then σ is a one-to-one mapping from $[N]$ onto itself.

II. PROBLEM STATEMENT

Consider a set of independent sources each producing i.i.d. data according to distinct but unknown probability distributions on a finite alphabet Z . Let $\mathcal{U} \subset \mathcal{P}(Z)$ denote the set of probability distributions followed by these sources. Let $\mathcal{M} \subseteq \mathcal{U}$ and $\mathcal{N} \subseteq \mathcal{U}$ be such that $\mathcal{M} \cap \mathcal{N} = \mathcal{K}$. Let $|\mathcal{M}| = M$, $|\mathcal{N}| = N$ and $|\mathcal{K}| = K$. We are concerned with the following two problems.

(P1) [Known sources] Let $\mathcal{M} = \{\mu_1, \mu_2, \dots, \mu_M\}$. Suppose M , N and K are known but \mathcal{N} and \mathcal{K} are not. Further, suppose a set $\mathcal{S} = \{y_1, y_2, \dots, y_N\}$ of unlabeled sequences of length n each generated independently under a distinct distribution in \mathcal{N} is given. Identify the K sequences in \mathcal{S} that were generated under distributions in \mathcal{M} , and match each of these sequences to the correct distribution in \mathcal{M} that generated it.

(P2) [*Unknown sources*] Suppose the distributions are unknown, but M , N and K are known. Given a set $\mathcal{S}_1 = \{x_1, x_2, \dots, x_M\}$ of unlabeled sequences of length n each generated under a distinct distribution in \mathcal{M} , and a set $\mathcal{S}_2 = \{y_1, y_2, \dots, y_N\}$ of unlabeled sequences of length n each generated under a distinct distribution in \mathcal{N} , identify the K sequences in \mathcal{S}_1 that were generated by distributions in \mathcal{K} and match each of them to the sequence in \mathcal{S}_2 that was generated under the same distribution. The information in \mathcal{S}_1 and \mathcal{S}_2 are assumed to be independent of each other.

As mentioned earlier, such problems arise in the fields of de-anonymization of databases, and of identification of users from the statistics of their data. For example, \mathcal{S}_1 and \mathcal{S}_2 in problem **(P2)** could be two anonymized databases of data belonging to known sets of users. It may be known that the two sets of users are identical, in which case $\mathcal{M} = \mathcal{N} = \mathcal{K}$, or it may be that the second set of users is a subset of the first set, in which case $\mathcal{M} \supset \mathcal{N} = \mathcal{K}$. In some other cases, the sets \mathcal{M} and \mathcal{N} might not be subsets but the statistician may have an estimate for the number K of common users in the two sets, i.e., the size of \mathcal{K} . Problem **(P1)** arises when the statistical behavior of the data belonging to the first set of users is known accurately. As stated above, for simplifying the analysis, in both these problems we have assumed that the sample size of all sequences are equal, and that the alphabet \mathcal{Z} is a finite set. In Section V we discuss how the analyses and results can be generalized to the setting in which the sequence lengths are not equal, or when the alphabet is continuous.

Both problems **(P1)** and **(P2)** can be visualized as variants of the following problem. Let $\mathcal{V}_1, \mathcal{V}_2$ be two sets of objects with $|\mathcal{V}_1| = M$ and $|\mathcal{V}_2| = N$. Consider a complete bipartite graph \mathcal{G} [13] with vertices $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ such that every vertex in \mathcal{V}_1 is connected to every vertex in \mathcal{V}_2 by an edge, as illustrated in Figure 3. The objective is to identify a matching¹ of cardinality K in the graph \mathcal{G} that satisfies some conditions. In problem **(P1)**, the set $\mathcal{V}_1 = \mathcal{M}$ and $\mathcal{V}_2 = \mathcal{S}$, and in problem **(P2)** the set $\mathcal{V}_1 = \mathcal{S}_1$ and $\mathcal{V}_2 = \mathcal{S}_2$. Illustrations of these matching problems are shown in Figure 2. More precisely, these problems are multi-hypothesis testing problems, where each hypothesis corresponds to a potential matching of cardinality K on the graph \mathcal{G} . Thus, for each problem, there are a total of $J = \binom{M}{K} \binom{N}{K} K!$ different hypotheses. We let $\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_J$ denote an enumeration of the hypotheses for each problem. We use the same notation for the hypotheses in both problems; it should always be clear from context what is intended. It is to be noted that in both problems the hypotheses are composite. In problem **(P1)**, the sequences in \mathcal{S} that do not follow a distribution in \mathcal{K} are allowed to have any distribution. In problem **(P2)** the probability distributions of each source could lie anywhere in $\mathcal{P}(\mathcal{Z})$.

¹A matching in a graph is a set of edges such that no two edges in the set share a common vertex.

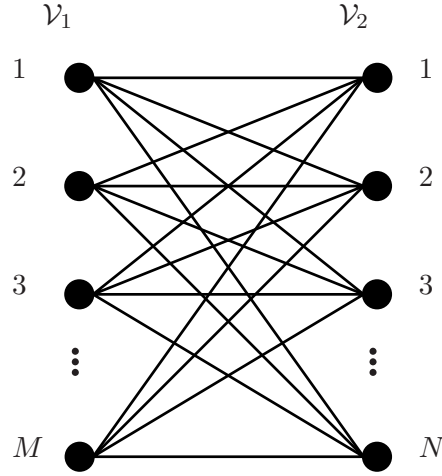


Fig. 3. A complete bipartite graph. Every vertex in \mathcal{V}_1 is connected by an edge to every vertex in \mathcal{V}_2 .

We seek decision rules for these problems that admit exponential decay of error probabilities as a function of n under each hypothesis. For this purpose, for each problem, we allow a no-match decision, i.e., rejection of all J hypotheses. Thus a decision rule for problem **(P1)** is given by a partition $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R)$ of $\mathbf{Z}^1 = (\mathbf{Z}^n)^N$ the space of vectors of the form y_1, y_2, \dots, y_N , into $(J + 1)$ disjoint cells $\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R$, where Ω_ℓ is the acceptance region for hypothesis \mathcal{H}_ℓ for $\ell \in [J]$, and $\Omega_R = \mathbf{Z}^1 - \cup_{\ell=1}^J \Omega_\ell$ is the rejection zone. Similarly, a decision rule for problem **(P2)** is given by a partition $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R)$ of $\mathbf{Z} = (\mathbf{Z}^n)^M \times (\mathbf{Z}^n)^N$ the space of vectors of the form $x_1, x_2, \dots, x_M, y_1, y_2, \dots, y_N$, into $(J + 1)$ disjoint cells $\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R$, where Ω_ℓ is the acceptance region for hypothesis \mathcal{H}_ℓ , and $\Omega_R = \mathbf{Z} - \cup_{\ell=1}^J \Omega_\ell$ is the rejection zone. In both these problems, we consider an error event err under hypothesis \mathcal{H}_ℓ to denote a decision in favor of a wrong hypothesis \mathcal{H}_k where $k \neq \ell$. We denote a decision in favor of rejection by rej . Note that a decision in favor of rejection does not correspond to an error event under any hypothesis. Thus, using the notation $\underline{x} = (x_1, x_2, \dots, x_M)$ and $\underline{y} = (y_1, y_2, \dots, y_N)$, the probability of error of the decision rule Ω under hypothesis \mathcal{H}_ℓ is given by

$$P_\Omega(\text{err}/\mathcal{H}_\ell) = \mathbb{P}_{\mathcal{H}_\ell} \left\{ \underline{y} \in \bigcup_{\substack{k=1 \\ k \neq \ell}}^J \Omega_k \right\} \quad (1)$$

for problem **(P1)**, and by

$$P_{\Omega}(\text{err}/\mathcal{H}_{\ell}) = P_{\mathcal{H}_{\ell}} \left\{ (\underline{x}, \underline{y}) \in \bigcup_{\substack{k=1 \\ k \neq \ell}}^J \Omega_k \right\} \quad (2)$$

for problem **(P2)**. Here $P_{\mathcal{H}_{\ell}}$ indicates the probability measure under hypothesis \mathcal{H}_{ℓ} . For both problems, we consider a generalized Neyman-Pearson criterion wherein we seek to ensure that all error probabilities decay exponentially in n with some predetermined slope λ , and simultaneously minimize the rejection probability subject to these constraints. Specifically, we seek optimal decision rules Ω such that $\forall \mathcal{U} \subset \mathcal{P}(\mathcal{Z})$

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Omega}(\text{err}/\mathcal{H}_{\ell}) \geq \lambda, \ell \in [J], \quad (3)$$

and Ω_R is minimal. The quantity on the left hand side of (3) is called the error exponent under hypothesis \mathcal{H}_{ℓ} . The optimality criterion for Problem **(P2)** is defined analogously. This approach for identifying an optimal test with rejection was introduced by Gutman in [4] when he solved Problem **(P2)** for $N = K = 1$.

We define rejection probabilities and rejection exponents analogously to error probability and error exponents. The probability of rejection of the decision rule Ω under hypothesis \mathcal{H}_{ℓ} is given by

$$P_{\Omega}(\text{rej}/\mathcal{H}_{\ell}) = P_{\mathcal{H}_{\ell}} \{ \underline{y} \in \Omega_R \} \quad (4)$$

for both problem **(P1)** and by

$$P_{\Omega}(\text{rej}/\mathcal{H}_{\ell}) = P_{\mathcal{H}_{\ell}} \{ (\underline{x}, \underline{y}) \in \Omega_R \} \quad (5)$$

for problem **(P2)**. The rejection exponents capture the rate of decay of the rejection probabilities and are defined as

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Omega}(\text{rej}/\mathcal{H}_{\ell}). \quad (6)$$

Some special cases of problem **(P1)** are listed below.

- 1) If $M \geq 2$ and $N = K = 1$, this is just the classical M -ary hypothesis testing problem, with rejection.
- 2) For general M, N, K , a variant of this hypothesis testing problem without rejection is discussed in [12]. The special case of $M = N = K = 2$ is considered in detail. In this case there are exactly two hypotheses. The authors solve the problem of optimizing one type of error exponent under a constraint on the other type of error exponent.

Specific versions of problem **(P2)** have been studied in the past. Some special cases of interest are listed below.

- 1) When $N = K = 1$, this is the problem studied by Gutman in [4]. The results and approach of the present paper are largely based on [4].
- 2) For $M = N = K$ we studied problem **(P2)** in a recent work [1].

In the present paper, we generalize these works to arbitrary choices of \mathcal{M} , \mathcal{N} and \mathcal{K} .

The main tools we use for proving the results in this paper are the *method of types* [14] and Sanov's theorem [15] (see also [16]). The following lemma (see e.g., [14, Ch. 11] for a proof) gives a bound on the probability of observing a sequence with a specific type, or equivalently, a specific empirical distribution.

Lemma II.1. *Let \mathcal{Y} be a finite set and $s \in \mathcal{Y}^n$ be an arbitrary string of length n with entries in \mathcal{Y} . Let $y \in \mathcal{Y}^n$ be a random string drawn i.i.d. under probability law $\nu \in \mathcal{P}(\mathcal{Y})$. Then*

$$2^{-n(D(\Gamma_s \parallel \nu) + \frac{|Z| \log(n+1)}{n})} \leq \mathbb{P}\{\Gamma_y = \Gamma_s\} \leq 2^{-nD(\Gamma_s \parallel \nu)} \quad (7)$$

where Γ_s and Γ_y represent the empirical distributions of s and y respectively. \square

Sanov's theorem is a statement on the behavior of the probability as $n \rightarrow \infty$. It characterizes the large deviations behavior of the empirical distribution of an i.i.d. sequence as stated below.

Theorem II.2 (Sanov [15]). *Let \mathcal{Y} be a finite set. For any $\nu \in \mathcal{P}(\mathcal{Y})$ if $y \in \mathcal{Y}^n$ is a random sequence of length n drawn i.i.d. under ν , and $A \subset \mathcal{P}(\mathcal{Y})$ such that A is the closure of its interior, then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}\{\Gamma_y \in A\} = -\min_{\mu \in A} D(\mu \parallel \nu).$$

\square

The main result of Section IV is based on the following lemma which gives a bound on large-deviations of a pair of empirical distributions.

Lemma II.3. *Let \mathcal{Y} be a finite set. For $i = 1, 2$ let $y_i \in \mathcal{Y}^n$ denote a length n string drawn i.i.d. under $\nu \in \mathcal{P}(\mathcal{Y})$. Further assume that y_1 and y_2 are mutually independent. Then we have*

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}\left\{\sum_{i=1}^2 D(\Gamma_{y_i} \parallel \frac{1}{2}(\Gamma_{y_1} + \Gamma_{y_2})) \geq \lambda\right\} \geq \lambda.$$

\square

We provide a proof in the appendix.

It is possible to generalize Sanov's theorem to the infinite alphabet setting in which \mathcal{Y} is countably or uncountably infinite (see, e.g., [16]). However, in this paper we focus only on the finite alphabet setting. The analysis of error probabilities and optimal tests in the continuous alphabet setting is much more involved, and are typically based on Cramer's theorem [16]. We present discussions on potential extensions of the results of this paper to other settings, including that of continuous alphabets, in Section V.

Before we present the solutions, we summarize the main results below.

- Optimal test for the matching problem **(P1)** with known source distributions, given in Theorem III.3.
- Comparison of error exponents and rejection exponents of the optimal test with the test that ignores constraints on sequences.
- Optimal test for the matching problem **(P2)** with unknown source distributions, given in Theorem IV.3.

III. OPTIMAL MATCHING WITH KNOWN DISTRIBUTIONS

In this section we solve problem **(P1)**. As described in Section II we use the optimality criterion based on error exponents given in (3). Let $\mathcal{M} = \{\mu_1, \mu_2, \dots, \mu_M\} \subset \mathcal{P}(\mathcal{Z})$ as before. Let \mathcal{G} be the graph in Figure 3 with \mathcal{V}_1 and \mathcal{V}_2 respectively representing \mathcal{M} and \mathcal{S} , both of which are described in the statement of problem **(P1)**. Let $M_\ell \subset \{1, 2, \dots, M\} \times \{1, 2, \dots, N\}$ with $|M_\ell| = K$ denote the matching on the complete bipartite graph \mathcal{G} under hypothesis \mathcal{H}_ℓ . Any edge $e \in M_\ell$ can be represented as $e = (e_1, e_2)$ with the understanding that the edge connects μ_{e_1} and y_{e_2} in graph \mathcal{G} . Thus, under hypothesis \mathcal{H}_ℓ we have

$$\mathcal{K} = \{\mu_{e_1} : e \in M_\ell\} = \mathcal{M} \cap \mathcal{N}$$

representing the probability distributions followed by the sources that produced the sequences in $\{y_{e_2} : e \in M_\ell\}$. There are $M - K$ sources in $\mathcal{M} \setminus \mathcal{K}$, which do not produce any sequence in \mathcal{S} , and there are $N - K$ sequences in \mathcal{S} that are produced by sources in $\mathcal{N} \setminus \mathcal{K}$.

Let

$$D(\mathcal{H}_\ell) = \sum_{(i,j) \in M_\ell} D(\Gamma_{y_j} \parallel \mu_i). \quad (8)$$

Consider the estimate for the hypothesis given by

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}(\underline{y}) = \arg \min_{\mathcal{H}_\ell} D(\mathcal{H}_\ell) \quad (9)$$

where the minimization is performed over all hypotheses. As each hypothesis is represented by a matching on \mathcal{G} with cardinality equal to K , the estimate of (9) can be interpreted as the hypothesis corresponding

to the minimum weight cardinality- K matching [13] on \mathcal{G} with appropriate weights assigned to the edges in \mathcal{G} . For $\mu_i \in \mathcal{M}$ and $y_j \in \mathcal{S}$ we let the weight w_{ij} of the edge between them to be

$$w_{ij} = D(\Gamma_{y_j} \parallel \mu_i). \quad (10)$$

Weight w_{ij} can be interpreted as a measure of the difference between distributions μ_i and Γ_{y_j} . Figure 4 shows the graph \mathcal{G} with weights added to the edges.

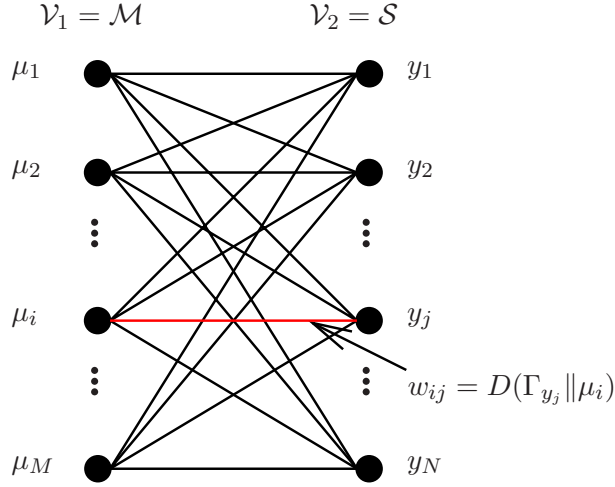


Fig. 4. The weighted complete bipartite graph \mathcal{G} for problem **(P1)**. The weight of the edge between the i -th vertex in \mathcal{V}_1 and the j -th vertex in \mathcal{V}_2 is given by (10). The matching corresponding to the hypothesis $\hat{\mathcal{H}}$ in (9) is given by the minimum weight matching on this graph with cardinality K .

We will now show that a test based on the estimate of $\hat{\mathcal{H}}$ in (9) is asymptotically optimal. For proving optimality we restrict ourselves to tests that are based only on the empirical distributions of the observations. Let Γ_Y denote the collection of empirical distributions:

$$\Gamma_Y := (\Gamma_{y_1}, \Gamma_{y_2}, \dots, \Gamma_{y_N}).$$

The restriction to tests based on empirical distributions is justified in the asymptotic setting because of the following lemma.

Lemma III.1. *Let $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R)$ be a decision rule based only on the distributions $\{\mu_1, \mu_2, \dots, \mu_M\}$ and $\{y_1, y_2, \dots, y_N\}$. Then there exists a decision rule $\Lambda = (\Lambda_1, \Lambda_2, \dots, \Lambda_J, \Lambda_R)$ based on the sufficient*

statistics Γ_Y such that

$$\begin{aligned}\liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_\Lambda(\text{err}/\mathcal{H}_\ell) &\geq \liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_\Omega(\text{err}/\mathcal{H}_\ell), \\ \liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_\Lambda(\text{rej}/\mathcal{H}_\ell) &\geq \liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_\Omega(\text{rej}/\mathcal{H}_\ell),\end{aligned}$$

for all $\ell \in [J]$ and for all choices of \mathcal{U} . \square

We provide a proof in the appendix. Thus this lemma suggests that if one is interested only in optimizing error exponents and rejection exponents, then tests based only on Γ_Y are sufficient.

Following Gutman [4], in order to prove optimality we allow for a no-match zone, i.e., we allow a decision in favor of rejecting all the M hypotheses. For this purpose, we need to identify the hypothesis corresponding to the second minimum weight matching in \mathcal{G} . Let

$$\tilde{\mathcal{H}} = \tilde{\mathcal{H}}(\underline{x}, \underline{y}) = \arg \min_{\mathcal{H}_\ell \neq \hat{\mathcal{H}}} D(\mathcal{H}_\ell) \quad (11)$$

where $\hat{\mathcal{H}}$ is defined in (9). The choices of $\hat{\mathcal{H}}$ and $\tilde{\mathcal{H}}$ have a simple interpretation in terms of maximum generalized likelihoods [2] as shown in the lemma below.

Lemma III.2. *The selections $\hat{\mathcal{H}}$ defined in (9) and $\tilde{\mathcal{H}}$ defined in (11) can be expressed as*

$$\hat{\mathcal{H}} = \mathcal{H}_{\hat{\ell}} \quad \text{and} \quad \tilde{\mathcal{H}} = \mathcal{H}_{\tilde{\ell}} \quad (12)$$

where

$$\begin{aligned}\hat{\ell} &= \arg \max_{\ell \in [J]} \max_{\substack{\mathcal{N} \subset \mathcal{P}(\mathcal{Z}) \\ \mathcal{M} \cap \mathcal{N} = \mathcal{K}}} P_{\mathcal{H}_\ell}(y_1, y_2, \dots, y_N) \\ \tilde{\ell} &= \arg \max_{\ell \in [J]: \mathcal{H}_\ell \neq \hat{\mathcal{H}}} \max_{\substack{\mathcal{N} \subset \mathcal{P}(\mathcal{Z}) \\ \mathcal{M} \cap \mathcal{N} = \mathcal{K}}} P_{\mathcal{H}_\ell}(y_1, y_2, \dots, y_N).\end{aligned}$$

\square

The above lemma is proved in the appendix. It is easy to see that in the special case that $M \geq N = K$, the set \mathcal{N} is fixed and thus the second minimization over the choice of distributions in \mathcal{N} is not necessary. In such a case the choice of $\hat{\mathcal{H}}$ can be interpreted as a simple maximum likelihood hypothesis. The optimal test with rejection can be described in terms of \hat{H} and \tilde{H} as shown in the following theorem.

Theorem III.3. *Let $\mathcal{M} = \{\mu_1, \mu_2, \dots, \mu_M\} \subset \mathcal{P}(\mathcal{Z})$ be a known set of M distinct probability distributions on the finite alphabet \mathcal{Z} . Let $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R)$ be a decision rule based on the collection Γ_Y of empirical distributions such that*

$$P_\Omega(\text{err}/\mathcal{H}_\ell) \leq 2^{-\lambda n}, \text{ for all } \ell \in [J] \quad (13)$$

and for all choices of distributions in $\mathcal{N} \setminus \mathcal{K}$.

$$\text{Let } \tilde{\lambda} = \lambda - \frac{N|\mathcal{Z}|\log(n+1)}{n},$$

$$\Lambda_\ell = \{\underline{y} : D(\tilde{\mathcal{H}}) \geq \tilde{\lambda}, \hat{\mathcal{H}} = \mathcal{H}_\ell\}, \ell \in [J],$$

and

$$\Lambda_R = \{\underline{y} : D(\tilde{\mathcal{H}}) < \tilde{\lambda}\}.$$

Then

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_\Lambda(\text{err}/\mathcal{H}_\ell) \geq \lambda, \quad \ell \in [J], \forall \mathcal{U} \subset \mathcal{P}(\mathcal{Z}) \quad (14)$$

and

$$\Lambda_R \subset \Omega_R. \quad (15)$$

□

We provide a proof to the theorem in the appendix. From the definition of $D(\mathcal{H}_\ell)$ in (8) it is clear that the decision regions Λ_ℓ 's and Λ_R proposed in Theorem III.3 depends on the sequences in \underline{y} only through Γ_Y . An illustration of the various decision regions of the optimal test as functions of Γ_Y is provided in Figure 5 for a specific example. From the condition of (15) it is obvious that the probability of rejection of the decision rules under the decision rule Λ is lower than the probability of rejection under the decision rule Ω . Thus the optimality result implies that the test Λ has lower rejection probabilities, as defined in (4), than any test Ω that satisfies an exponential decay of error probabilities as in (13).

The test can be explained in words as follows. First identify the hypotheses corresponding to the minimum weight matching and the second minimum weight matching of cardinality K in \mathcal{G} . Accept the former hypothesis if the weight corresponding to the latter exceeds the threshold $\tilde{\lambda}$, and reject all hypotheses if the threshold is not exceeded. When $M \geq N = K$, the result of Lemma III.2 implies that this test leads to a rejection if the weights corresponding to the two most likely hypotheses, $\hat{\mathcal{H}}$ and $\tilde{\mathcal{H}}$, are below a threshold, or equivalently, if the observations can be well-explained by two or more hypotheses.

We note that the threshold $\tilde{\lambda}$ appearing in the definition of Λ_R satisfies $\tilde{\lambda} \rightarrow \lambda$ as $n \rightarrow \infty$. Using Sanov's theorem we show in the proof that the choice of decision regions ensures that the error-exponent constraint of (14) is satisfied. We also observe that the offset between $\tilde{\lambda}$ and λ is just N times the offset in the exponent appearing in the first inequality of (7) which bounds the probability of observing a type. This offset is introduced to ensure that the condition (15) is satisfied, as detailed in the proof.

$$(\mathcal{P}(Z))^2$$

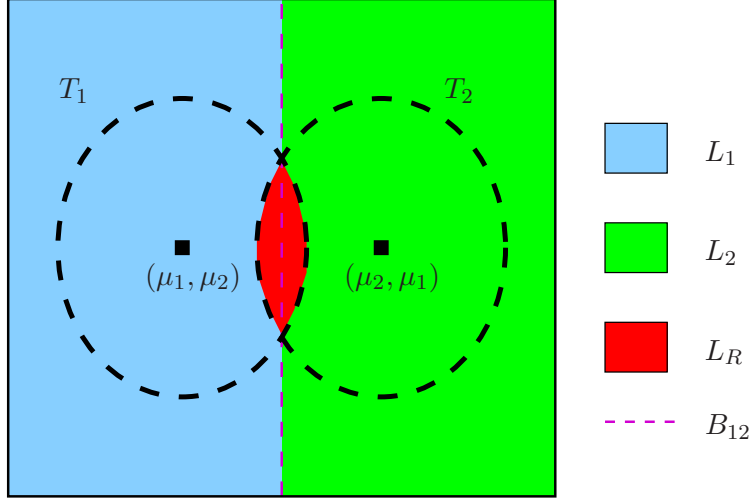


Fig. 5. Illustration of the decision regions for the optimal test of Theorem III.3 for $M = N = K = 2$, which means that $J = 2$. Here we define $T_\ell = \{(\nu_1, \nu_2) : \sum_{(i,j) \in M_\ell} D(\nu_j \| \mu_i) < \lambda\}$. Assuming M_1 denotes the matching in which y_i is matched to μ_i , and M_2 denotes the matching in which y_1 is matched to μ_2 and y_2 to μ_1 , it follows that $(\mu_1, \mu_2) \in T_1$ and $(\mu_2, \mu_1) \in T_2$. We use B_{12} to denote the hyperplane separating L_1 and L_2 defined as $B_{12} = \{(\nu_1, \nu_2) : \sum_{(i,j) \in M_1} D(\nu_j \| \mu_i) = \sum_{(i,j) \in M_2} D(\nu_j \| \mu_i)\}$, or equivalently, $B_{12} = \{(\nu_1, \nu_2) : \sum_{z \in Z} (\nu_1(z) - \nu_2(z)) \log \frac{\mu_1(z)}{\mu_2(z)} = 0\}$. Then the optimal decision regions of Theorem III.3 can be expressed as $\Lambda_i = \{\underline{y} : \Gamma_Y \in L_i\}$ for $i \in \{1, 2, R\}$, where L_i are as shown in the figure. Furthermore, if Γ_Y lies to the left of B_{12} in the figure, then $\hat{\mathcal{H}} = \mathcal{H}_1$ and if Γ_Y lies to the right of B_{12} then $\hat{\mathcal{H}} = \mathcal{H}_2$.

A. Comparison with the unconstrained problem

As we mentioned earlier, the problem studied in this section differs from ordinary multiple hypothesis testing because of the prior knowledge that the strings in \mathcal{S} were generated by distinct sources. It is interesting to compare the performance obtained by using the optimal test that makes use of this information with the performance of the optimal test that one would have to use in the absence of this prior information. Before we proceed we need to introduce some new notations. Let $\pi, \pi_1, \pi_2 \in \mathcal{P}(\mathcal{Y})$ be distinct probability mass functions with complete supports on \mathcal{Y} . For any $\eta \geq 0$ we define

$$Q_\eta(\pi) := \{\nu \in \mathcal{P}(\mathcal{Y}) : D(\nu \| \pi) < \eta\}, \quad (16)$$

and

$$E_\eta(\pi_1, \pi_2) = \sup\{\beta \geq 0 : Q_\beta(\pi_2) \cap Q_\eta(\pi_1) = \emptyset\}. \quad (17)$$

The function $E_\eta(\pi_1, \pi_2)$ is strictly monotonically decreasing in η in the interval $\eta \in (0, D(\pi_2 \| \pi_1))$ (see, e.g., [14, Sec. 11.7], [17, Sec. 3.2], and [18]). Moreover, if $\eta \geq D(\pi_2 \| \pi_1)$, then $E_\eta(\pi_1, \pi_2) = 0$. The

Chernoff information [14] between π_1 and π_2 is defined as

$$C(\pi_1, \pi_2) := - \inf_{\alpha \in [0,1]} \log \left(\sum_{y \in \mathcal{Y}} \pi_1^\alpha(y) \pi_2^{1-\alpha}(y) \right).$$

It is well known [14], [17] that

$$E_{C(\pi_1, \pi_2)}(\pi_1, \pi_2) = C(\pi_1, \pi_2).$$

These quantities are illustrated in Figure 6.

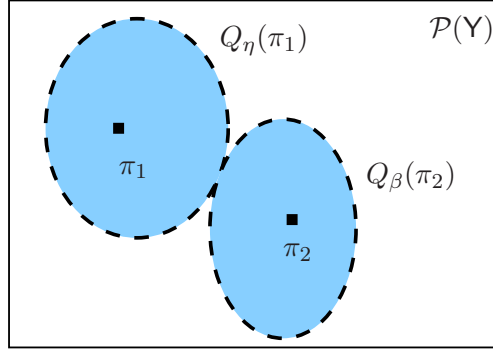


Fig. 6. Illustration of the Q_η sets defined in (16) on the probability simplex $\mathcal{P}(\mathcal{Y})$. In the above figure, the sets $Q_\eta(\pi_1)$ and $Q_\beta(\pi_2)$ touch each other and hence $\beta = E_\eta(\pi_1, \pi_2)$ as defined in (17). If in addition $\beta = \eta$ then $\eta = C(\pi_1, \pi_2)$.

Below we provide two comparisons, the first being a comparison of the rejection regions in the two cases, and the second, a comparison of the error exponents obtained by tests that do not allow for a rejection region.

1) *Comparison of rejection regions:* In the absence of prior knowledge that the strings in \mathcal{S} were generated by distinct sources, one is forced to repeatedly perform optimal multihypothesis testing on each string in \mathcal{S} . In other words, for each string s in \mathcal{S} , one repeats the optimal solution of Theorem III.3 assuming that the second set \mathcal{S} is a singleton comprising the single string s . These individual solutions can then be combined to obtain a solution to the original problem as follows.

Assume $M \geq N = K$. Consider the function $\hat{\sigma} : [N] \mapsto [M]$ defined by

$$\hat{\sigma}(i) = \arg \min_{j \in [M]} D(\Gamma_{y_i} \| \mu_j), \text{ for all } i \in [N]. \quad (18)$$

Thus the function $\hat{\sigma}$ gives the best matching of each string to one of the sources in \mathcal{M} . In fact, it can be shown that $\hat{\sigma}(i)$ is the maximum likelihood source that produced y_i , just as in Lemma III.2. If $\hat{\sigma}$ is a one-to-one function, then it corresponds to a valid hypothesis for the matching problem. We call this hypothesis $\mathcal{H}^{\hat{\sigma}}$. If $\hat{\sigma}$ is not a one-to-one function, or if $\mathcal{H}^{\hat{\sigma}}$ does not correspond to the true hypothesis,

then the strings are not correctly matched and hence in this case an error occurs. Furthermore, in order to satisfy the error exponent constraint, one is forced to reject whenever the individual hypothesis test on any of the N strings leads to a rejection. Let

$$\tilde{w}_i = \min_{j \in [M] \setminus \hat{\sigma}(i)} D(\Gamma_{y_i} \parallel \mu_j), \quad i \in [N]$$

The solution to the overall problem is now given by

$$\Lambda_\ell^{\text{uc}} = \{\underline{y} : \min_{i \in [N]} \tilde{w}_i \geq \check{\lambda}, \mathcal{H}^{\hat{\sigma}} = \mathcal{H}_\ell\}, \ell \in [J] \quad (19)$$

where the superscript of uc indicates that the solution is unconstrained and

$$\Lambda_R^{\text{uc}} = \{\underline{y} : \min_{i \in [N]} \tilde{w}_i < \check{\lambda}\} \quad (20)$$

where $\check{\lambda} = \lambda - \frac{|Z| \log(n+1)}{n}$, is the optimal choice for the threshold obtained from Theorem III.3 when the set of strings is a singleton. The probability of error of this solution is given by

$$\begin{aligned} P_{\Lambda^{\text{uc}}}(\text{err} | \mathcal{H}_\ell) &= \mathbb{P}_{\mathcal{H}_\ell} \{y_i \text{ is incorrectly matched for some } i \in [N]\} \\ &\leq \sum_{i=1}^N \mathbb{P}_{\mathcal{H}_\ell} \{y_i \text{ is incorrectly matched}\} \end{aligned}$$

where the inequality follows via the union bound. By the result of Theorem III.3, each term in the above summation decays exponentially in n with exponent λ and thus we have

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Lambda^{\text{uc}}}(\text{err} | \mathcal{H}_\ell) \geq \lambda.$$

Thus this solution meets the same error exponent constraint as the optimal solution of Theorem III.3 that one can use when the constraint on the strings is known a priori. However, the rejection regions for the optimal test is a strict subset of the rejection region of (20), as is evident from the conclusion of Theorem III.3. For large n , we have $\tilde{\lambda} \approx \check{\lambda} \approx \lambda$, and thus the sizes of the rejection regions can be significantly different. The significance can be quantified by comparing the probability of rejection of the two tests. For large n , it is easier to look at the large deviations behavior of these probabilities for which we use rejection exponents. To keep the presentation simple, in the rest of this section, we focus on the setting in which $M = N = K$. Furthermore, we assume that the probability distributions in \mathcal{M} are distinct.

Let $\sigma^i \in \text{Sym}([N]), i \in [J]$ where $J = N!$ denote an enumeration of all possible bijections from $[N]$ onto itself, i.e.,

$$\sigma^i : [N] \mapsto [N], \quad i \in [J]$$

represents a unique permutation of $[N]$ for each i . For each i let μ^{σ^i} denote the product distribution $\mu_{\sigma^i(1)} \times \mu_{\sigma^i(2)} \times \dots \mu_{\sigma^i(N)}$. It follows, via a straightforward application of Sanov's theorem that the rejection exponents of the optimal test given in Theorem III.3 and the test Λ^{uc} given in (19) and (20) can be expressed as follows:

$$\begin{aligned} & \liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Lambda}(\text{rej}/\mathcal{H}_{\ell}) \\ &= \begin{cases} \min_{\substack{i,j \in [J] \\ i \neq j}} E_{\lambda}(\mu^{\sigma^i}, \mu^{\sigma^j}) & \text{if } C^* < \lambda \\ \infty & \text{else} \end{cases} \end{aligned} \quad (21)$$

where $C^* = \min_{\substack{i,j \in [J] \\ i \neq j}} C(\mu^{\sigma^i}, \mu^{\sigma^j})$, and

$$\begin{aligned} & \liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Lambda^{\text{uc}}}(\text{rej}/\mathcal{H}_{\ell}) \\ &= \begin{cases} \min_{\substack{i,j \in [N] \\ i \neq j}} E_{\lambda}(\mu_i, \mu_j) & \text{if } C^{\text{uc}*} < \lambda \\ \infty & \text{else} \end{cases} \end{aligned} \quad (22)$$

where $C^{\text{uc}*} = \min_{\substack{i,j \in [N] \\ i \neq j}} C(\mu_i, \mu_j)$. Thus the important quantities that determine the rejection exponent in the former case are the minimum value of the E_{λ} function and Chernoff information measured between pairs of μ^{σ^i} distributions, and in the latter case, the same functions measured between pairs of μ_i distributions. These quantities can differ significantly as we illustrate in Example III.1 later in the paper.

2) *Comparison of error probabilities without rejection:* An alternative version of the problem studied in this section is to try to identify an optimal test that does not allow rejection as a test outcome. When $M \geq N = K$, the problem studied here is just a standard multihypothesis testing problem with J different hypotheses, one corresponding to each permutation of N distributions from the set $\{\mu_1, \mu_2, \dots, \mu_M\}$. In this setting, the solution $\hat{\mathcal{H}}$ given by (9) is in fact the maximum-likelihood solution as shown in (12) in Lemma III.2. Also, by applying Lemma III.2 to the setting in which $N = 1$, it follows that the solution $\hat{\sigma}$ of (18) can be expressed as

$$\hat{\sigma}(i) = \arg \max_{j \in [M]} \mu_j(y_i). \quad (23)$$

Thus the solution $\hat{\sigma}$ of (18) is the maximum likelihood (ML) solution for the problem studied in this paper when the constraint on the strings is unknown, i.e., it is the ML solution when each string has to be independently classified to one of the sources without any constraints. For classical multihypothesis testing problems without rejection, the maximum likelihood solution is known to be asymptotically optimal in

terms of maximizing the worst-case error exponent [19] among all hypotheses. Furthermore, the value of the error exponent is given by the Chernoff information [19]. In fact it is straightforward to show that

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\mathcal{H}_\ell} \{\widehat{\mathcal{H}} \neq \mathcal{H}_\ell\} = C^*, \text{ for all } \ell \in [J]. \quad (24)$$

Furthermore, if $\sigma^\ell \in \text{Sym}([N])$ denotes the permutation function such that y_i is drawn from source $\mu_{\sigma^\ell(i)}$ under hypothesis \mathcal{H}_ℓ , then

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\mathcal{H}_\ell} \bigcup_{i=1}^N \{\widehat{\sigma}(i) \neq \sigma^\ell(i)\} = C^{\text{uc}*}, \text{ for all } \ell \in [J] \quad (25)$$

where $\widehat{\sigma}$ is given by (18).

Comparing with (21) and (22) we see that the the Chernoff informations C^* and $C^{\text{uc}*}$ which determine the error exponents for these tests are equal to the critical values of the error exponent constraints λ in the test with rejection, below which the rejection exponent is ∞ . In the following example we show that the Chernoff informations C^* and $C^{\text{uc}*}$ for the constrained and unconstrained problems can be significantly different. Thus the optimal error exponents and rejection exponents in the constrained setting can be significantly higher than those in the unconstrained setting.

Example III.1. As a simple example, suppose $M = N = K = 2$, and $\mathcal{Z} = \{0, 1\}$. Let μ_1 be given by the Bernoulli distribution with parameter $\frac{1}{2}$ and μ_2 a Bernoulli distribution with parameter ρ . In this case $J = N! = 2$ and the two possible permutations are σ^1 and σ^2 where σ^1 is the identity function on $\{1, 2\}$ and

$$\sigma^2(1) = 2 \quad \text{and} \quad \sigma^2(2) = 1.$$

In this case the distributions $\mu^{\sigma^1} = \mu_1 \times \mu_2$ and $\mu^{\sigma^2} = \mu_2 \times \mu_1$. These distributions are illustrated in Table I.

TABLE I
PROBABILITY MASS FUNCTIONS ILLUSTRATED

(a) PMFs μ_1 and μ_2			(b) PMFs μ^{σ^1} and μ^{σ^2}				
	0	1		00	01	10	11
μ_1	$\frac{1}{2}$	$\frac{1}{2}$	μ^{σ^1}	$\frac{1}{2}(1-\rho)$	$\frac{1}{2}(\rho)$	$\frac{1}{2}(1-\rho)$	$\frac{1}{2}(\rho)$
μ_2	$1-\rho$	ρ	μ^{σ^2}	$(1-\rho)\frac{1}{2}$	$(1-\rho)\frac{1}{2}$	$(\rho)\frac{1}{2}$	$(\rho)\frac{1}{2}$

According to the definitions, in this case, the Chernoff informations are given by

$$C^* = C(\mu^{\sigma^1}, \mu^{\sigma^2}) \quad \text{and} \quad C^{\text{uc}*} = C(\mu_1, \mu_2).$$

These quantities are illustrated as a function of ρ in Figure 7. As we see in the figures these quantities

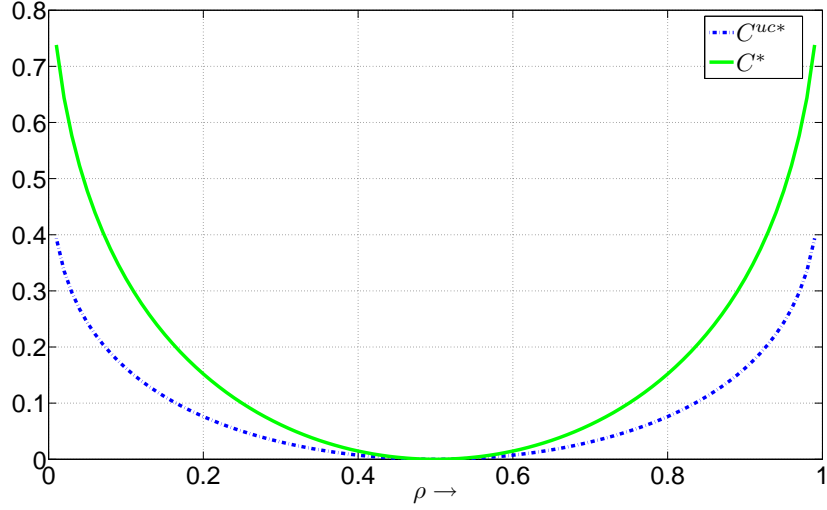


Fig. 7. Comparison of Chernoff informations C^{uc*} and C^* for a simple example with $M = N = K = 2$ and μ_1 given by the Bernoulli distribution with parameter $\frac{1}{2}$ and μ_2 the Bernoulli distribution with parameter ρ .

are different for all $\rho \neq \frac{1}{2}$ and the difference can be significant.

□

Thus we conclude from Example III.1 and the results of (21), (22), (24), and (25), that by using the unconstrained solution rather than the constrained solution we get a performance improvement in terms of the error exponent. However, the unconstrained solution has a practical advantage over the optimal solution in terms of the computational complexity of the algorithm for determining the solution, as we elaborate in Section V.

IV. OPTIMAL MATCHING WITH UNKNOWN DISTRIBUTIONS

In this section we solve problem **(P2)**. The structure of the solution is very similar to that we obtained in Section III. As before we use the optimality criterion based on error exponents given in (3). In this section we use \mathcal{G} to denote the graph in Figure 3 with \mathcal{V}_1 and \mathcal{V}_2 respectively representing \mathcal{S}_1 and \mathcal{S}_2 , both of which are described in the statement of problem **(P2)**. Let $M_\ell \subset \{1, 2, \dots, M\} \times \{1, 2, \dots, N\}$ with $|M_\ell| = K$ denote the matching on \mathcal{G} under hypothesis \mathcal{H}_ℓ . Analogous to our notation in Section III, edge $e \in M_\ell$ can be represented as $e = (e_1, e_2)$ with the understanding that the edge connects x_{e_1} and y_{e_2} .

in graph \mathcal{G} . Recall that $\mathcal{M}(\mathcal{N})$ represents the probability distributions followed by the $M(N)$ sources that produced the sequences in $\mathcal{S}_1(\mathcal{S}_2)$, and that $\mathcal{M} \cap \mathcal{N} = \mathcal{K}$ with $|\mathcal{K}| = K$. Thus there are $M - K$ sequences in \mathcal{S}_1 and $N - K$ sequences in \mathcal{S}_2 that are not produced by sources in \mathcal{K} .

Let

$$D(\mathcal{H}_\ell) = \sum_{(i,j) \in \mathcal{M}_\ell} D(\Gamma_{x_i} \parallel \frac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})) + D(\Gamma_{y_j} \parallel \frac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})). \quad (26)$$

Consider the estimate for the hypothesis given by

$$\hat{\mathcal{H}} = \arg \min_{\mathcal{H}_\ell} D(\mathcal{H}_\ell) \quad (27)$$

This test can be interpreted as a minimum weight cardinality- K matching [13] on the complete bipartite graph \mathcal{G} with appropriate weights assigned to the edges in \mathcal{G} . For $x_i \in \mathcal{S}_1$ and $y_j \in \mathcal{S}_2$ let the weight w_{ij} of the edge between them be given by

$$w_{ij} = D(\Gamma_{x_i} \parallel \frac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})) + D(\Gamma_{y_j} \parallel \frac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})). \quad (28)$$

As the sequences x_i and y_j have equal lengths, the quantity $\frac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})$ appearing in (28) can be interpreted as the empirical distribution of the concatenation of x_i and y_j . Thus weight w_{ij} can be interpreted as the sum of two quantities – the first quantity representing a measure of the difference between sequence x_i and the concatenated sequence, and the second quantity representing a measure of the difference between y_j and the concatenated sequence. Effectively, w_{ij} can be interpreted as a different distance measure between sequences x_i and y_j . Figure 8 shows the graph \mathcal{G} with weights added to the edges.

We will now show that a test based on the estimate of $\hat{\mathcal{H}}$ in (27) is asymptotically optimal. For proving asymptotic optimality we restrict ourselves to tests that are based only on the empirical distributions of the observations. Let Γ_{XY} denote the collection of empirical distributions:

$$\Gamma_{XY} := (\Gamma_{x_1}, \Gamma_{x_2}, \dots, \Gamma_{x_M}, \Gamma_{y_1}, \Gamma_{y_2}, \dots, \Gamma_{y_N}).$$

The restriction to tests based on empirical distributions is justified in the asymptotic setting because of the following lemma.

Lemma IV.1. *Let $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R)$ be a decision rule based only on the sequences $\{x_1, x_2, \dots, x_M\}$ and $\{y_1, y_2, \dots, y_N\}$. Then there exists a decision rule $\Lambda = (\Lambda_1, \Lambda_2, \dots, \Lambda_J, \Lambda_R)$ based on the sufficient*

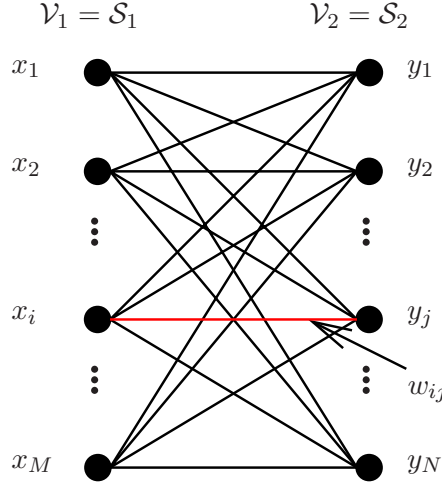


Fig. 8. The weighted complete bipartite graph \mathcal{G} for problem **(P2)**. The weight of the edge between the i -th vertex in \mathcal{V}_1 and the j -th vertex in \mathcal{V}_2 is given by (28). The matching corresponding to the hypothesis $\hat{\mathcal{H}}$ in (27) is given by the minimum weight matching on this graph with cardinality- K .

statistics Γ_{XY} such that

$$\begin{aligned} \liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Lambda}(\text{err}/\mathcal{H}_{\ell}) &\geq \liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Omega}(\text{err}/\mathcal{H}_{\ell}), \\ \liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Lambda}(\text{rej}/\mathcal{H}_{\ell}) &\geq \liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Omega}(\text{rej}/\mathcal{H}_{\ell}). \end{aligned}$$

for all $\ell \in [J]$ and all $\mathcal{U} \subset \mathcal{P}(\mathcal{Z})$ □

We provide a proof in the appendix. Note that Λ_R and Ω_R are finite sets, thus their cardinality is well-defined.

In order to ensure exponential decay of the error probabilities at a prescribed exponential rate, we allow a no-match zone, i.e., we allow a decision in favor of rejecting all the M hypotheses. For this purpose, we need to identify the hypothesis corresponding to the second minimum weight matching in \mathcal{G} . Let

$$\tilde{\mathcal{H}} = \arg \min_{\mathcal{H}_{\ell} \neq \hat{\mathcal{H}}} D(\mathcal{H}_{\ell}) \quad (29)$$

where $\hat{\mathcal{H}}$ is defined in (27). As in Section III, the choices of $\hat{\mathcal{H}}$ and $\tilde{\mathcal{H}}$ have a simple interpretation based on maximum *generalized likelihoods* as was the case in Lemma III.2.

Lemma IV.2. *The selections $\hat{\mathcal{H}}$ defined in (27) and $\tilde{\mathcal{H}}$ defined in (29) can be expressed as*

$$\hat{\mathcal{H}} = \arg \max_{\ell \in [J]} \max_{\substack{\mathcal{M}, \mathcal{N} \subset \mathcal{P}(\mathcal{Z}) \\ |\mathcal{M} \cap \mathcal{N}| = K}} \mathbb{P}_{\mathcal{H}_\ell}(x_1, \dots, x_N, y_1, \dots, y_N), \quad (30)$$

$$\tilde{\mathcal{H}} = \arg \max_{\ell \in [J]: \mathcal{H}_\ell \neq \hat{\mathcal{H}}} \max_{\substack{\mathcal{M}, \mathcal{N} \subset \mathcal{P}(\mathcal{Z}) \\ |\mathcal{M} \cap \mathcal{N}| = K}} \mathbb{P}_{\mathcal{H}_\ell}(x_1, \dots, x_N, y_1, \dots, y_N). \quad (31)$$

□

The above lemma is proved in the appendix. As in Section III, the optimal test with rejection can be stated in terms of $\hat{\mathcal{H}}$ and $\tilde{\mathcal{H}}$ as described in the following theorem.

Theorem IV.3. *Let $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R)$ be a decision rule based on the collection Γ_{XY} of empirical distributions such that*

$$P_\Omega(\text{err}/\mathcal{H}_\ell) \leq 2^{-\lambda n}, \text{ for all } \ell \in [J] \quad (32)$$

and for all choices of distributions in \mathcal{K} , $\mathcal{M} \setminus \mathcal{K}$, and $\mathcal{N} \setminus \mathcal{K}$.

$$\text{Let } \tilde{\lambda} = \lambda - \frac{(M+N)|\mathcal{Z}| \log(n+1)}{n},$$

$$\Lambda_\ell = \{(\underline{x}, \underline{y}) : D(\tilde{\mathcal{H}}) \geq \tilde{\lambda}, \hat{\mathcal{H}} = \mathcal{H}_\ell\}, \ell \in [J],$$

and

$$\Lambda_R = \{(\underline{x}, \underline{y}) : D(\tilde{\mathcal{H}}) < \tilde{\lambda}\}.$$

Then

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_\Lambda(\text{err}/\mathcal{H}_\ell) \geq \lambda, \quad \ell \in [J], \forall \mathcal{U} \subset \mathcal{P}(\mathcal{Z}) \quad (33)$$

and

$$\Lambda_R \subset \Omega_R. \quad (34)$$

□

We provide a proof to the theorem in the appendix. From the definition of $D(\mathcal{H}_\ell)$ in (26) it is clear that the decision regions Λ_ℓ 's and Λ_R proposed in Theorem IV.3 depends on the sequences in \underline{x} and \underline{y} only through Γ_{XY} . From the condition of (34) it is obvious that the probability of rejection of the decision rules under the decision rule Λ is lower than the probability of rejection under the decision rule Ω . Thus the optimality result implies that the test Λ has lower rejection probabilities, as defined in (5),

than any test Ω that satisfies an exponential decay of error probabilities as in (32). For $N = 1$, this result is similar to that given by Gutman in [4, Thm 2]. However, the rejection condition in this solution is different from that provided by Gutman and can be interpreted as the condition under which the second lowest weight matching has a weight below a threshold.

The test can be explained in words as follows. First identify the hypothesis corresponding to the minimum weight matching and the second minimum weight matching of cardinality K in \mathcal{G} . Accept the former hypothesis if the weight corresponding to the latter exceeds the threshold $\tilde{\lambda}$, and reject all hypotheses if the threshold is not exceeded. In the case of $M = N = K$, it follows via Lemma IV.2 that the optimal choice of the hypothesis is given by the maximum generalized likelihood hypothesis, and that a no-match decision is selected when the second highest generalized likelihood exceeds a threshold. In other words, this test leads to a rejection if the observations can be well-explained by two or more hypotheses.

We note that the threshold $\tilde{\lambda}$ appearing in the definition of Λ_R satisfies $\tilde{\lambda} \rightarrow \lambda$ as $n \rightarrow \infty$. Using a generalization of Lemma II.3 we show in the proof that the choice of decision regions ensures that the error-exponent constraint of (33) is satisfied. We also observe that the offset between $\tilde{\lambda}$ and λ is just $(M + N)$ times the offset in the exponent appearing in the first inequality of (7) which bounds the probability of observing a type. This offset is introduced to ensure that the condition (34) is satisfied. The detailed arguments are in the proof.

A. Comparison with the unconstrained problem

As in Section III, it is interesting to compare the solution of Theorem IV.3 with the solution that one would have to use without the prior knowledge that the strings in \mathcal{S}_2 (also \mathcal{S}_1) are produced by distinct sources. We follow the same steps as in Section III-A. Assume $M = N = K$. In the absence of prior knowledge, a reasonable strategy is to try to sequentially match each string in \mathcal{S}_2 to some string in \mathcal{S}_1 . For each $i \in [N]$ define

$$\hat{\sigma}(i) = \arg \min_{j \in [M]} D(\Gamma_{x_j} \| \frac{1}{2}(\Gamma_{x_j} + \Gamma_{y_i})) + D(\Gamma_{y_i} \| \frac{1}{2}(\Gamma_{x_j} + \Gamma_{y_i})). \quad (35)$$

The function $\hat{\sigma}$ maps every string in \mathcal{S}_2 to some string in \mathcal{S}_1 . If $\hat{\sigma}$ is a one-to-one function, then it corresponds to a valid hypothesis for the matching problem. We call this hypothesis $\mathcal{H}^{\hat{\sigma}}$. If $\hat{\sigma}$ is not a one-to-one function, or if $\mathcal{H}^{\hat{\sigma}}$ does not correspond to the true hypothesis, then the strings are not correctly matched and hence in this case an error occurs. Furthermore, in order to satisfy the error exponent constraint, one is forced to reject whenever the individual hypothesis test on any of the N

strings in \mathcal{S}_2 leads to a rejection. For $i \in [N]$, let

$$\tilde{w}_i = \min_{j \in [M] \setminus \hat{\sigma}(i)} D(\Gamma_{x_j} \| \frac{1}{2}(\Gamma_{x_j} + \Gamma_{y_i})) + D(\Gamma_{y_i} \| \frac{1}{2}(\Gamma_{x_j} + \Gamma_{y_i})).$$

The solution to the overall problem is now given by

$$\Lambda_\ell^{\text{uc}} = \{(\underline{x}, \underline{y}) : \min_{i \in [N]} \tilde{w}_i \geq \check{\lambda}, \mathcal{H}^{\hat{\sigma}} = \mathcal{H}_\ell\}, \ell \in [J] \quad (36)$$

where the superscript of uc indicates that the solution is unconstrained and

$$\Lambda_R^{\text{uc}} = \{(\underline{x}, \underline{y}) : \min_{i \in [N]} \tilde{w}_i < \check{\lambda}\} \quad (37)$$

where $\check{\lambda} = \lambda - \frac{(N+1)|Z|\log(n+1)}{n}$, is the optimal choice for the threshold obtained from Theorem IV.3 when the second set of strings \mathcal{S}_2 is a singleton. The probability of error of this solution is given by

$$\begin{aligned} P_{\Lambda^{\text{uc}}}(\text{err}|\mathcal{H}_\ell) &= \mathbb{P}_{\mathcal{H}_\ell}\{y_i \text{ is incorrectly matched for some } i \in [N]\} \\ &\leq \sum_{i=1}^N \mathbb{P}_{\mathcal{H}_\ell}\{y_i \text{ is incorrectly matched}\} \end{aligned}$$

where the inequality follows via the union bound. By the result of Theorem IV.3, each term in the above summation decays exponentially in n with exponent λ and thus we have

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log P_{\Lambda^{\text{uc}}}(\text{err}|\mathcal{H}_\ell) \geq \lambda.$$

Thus this solution meets the same error exponent constraint as the optimal solution of Theorem IV.3 that one can use when the constraint on the strings is known a priori. However, the rejection regions for the optimal test is a strict subset of the rejection region of (37), as is evident from the conclusion of Theorem IV.3. These regions can be visualized using the graph \mathcal{G} of Figure 8. The rejection region Λ_R of the optimal test of Theorem IV.3 corresponds to all sequences $(\underline{x}, \underline{y})$ such that there are two cardinality K matchings on \mathcal{G} with weight less than or equal to $\check{\lambda}$ where the weight of a matching is given by (8). However, the rejection region Λ_R^{uc} of the unconstrained solution is the set of all $(\underline{x}, \underline{y})$ such that some node on the right hand partition has two edges with weight less than or equal to $\check{\lambda}$. For large n , we have $\check{\lambda} \approx \check{\lambda} \approx \lambda$, and thus the sizes of the rejection regions can be significantly different. As in Section III-A1, it is possible to use Sanov's theorem to quantify the significance by comparing the rejection exponents of the two tests. Similarly it is possible to compare error exponents obtained by tests that do not allow rejection. However, the analysis is much more involved as in the problem **(P2)** we now have two strings from each source, and hence analytical expressions for the rejection exponents are difficult to obtain. We therefore avoid the details in this paper.

V. PRACTICAL ASPECTS, EXTENSIONS AND CONCLUSIONS

We proposed asymptotically optimal solutions to two hypothesis testing problems that seek to match unlabeled sequences of observations to labeled source distributions or training sequences. Under the constraint that the observed sequences are drawn from distinct sources, the structure of the optimal solution is significantly different from the unconstrained solution, and can lead to significant improvement in performance as we saw in Sections III-A and IV-A.

An important practical aspect is that of the complexity of the algorithms required to identify the optimal solutions of Theorem III.3 and IV.3. The unconstrained solutions of (18) and (35) are straightforward to identify because these can be obtained by sequentially matching each string in \mathcal{S} or \mathcal{S}_2 to one of the M sources in \mathcal{M} or one of the M strings in \mathcal{S}_1 . This leads to a time-complexity of $O(MN)$. The optimal (constrained) solutions are in general more complex to identify as a combinatorial optimization problem has to be solved to identify \hat{H} and \tilde{H} defined in (9), (27), (11) and (29). Nevertheless, these solutions can be identified by solving minimum weight bipartite matching problems on the graphs \mathcal{G} which can be executed efficiently in polynomial-time.

The first step in implementing the solutions to problems **(P1)** and **(P2)** is to identify the estimates of (9) and (27). As discussed earlier, the task of identifying these estimates is equivalent to solving a minimum weight cardinality- K matching problem on a weighted complete bipartite graph. If $M = N = K$, then this problem can be solved using the Hungarian algorithm [20], which has a time-complexity of $O(N^3)$ (see [21] and references therein). When $M \neq N$, the Hungarian algorithm can be adapted to run in $O(MNK)$ as detailed in [22]. Thus the complexity of this algorithm is roughly K times more than that of the naive unconstrained algorithm. This problem can also be solved using a polynomial time algorithm based on the theory of matroids (see, e.g., [23, Ch. 8]). In practice the complexity can often be reduced significantly. For instance, in the solution of (9), if some empirical distribution Γ_{y_j} is not absolutely continuous with respect to some μ_i , then the edge connecting the corresponding vertices in \mathcal{G} can be removed, as it will never be selected in the minimum weight matching. The same step can be performed if the empirical distributions Γ_{x_i} and Γ_{y_j} have disjoint supports in the solution of (27). If the number of remaining edges in the graph is E , then, when $M = N = K$, the Hungarian algorithm can be adapted to run with a complexity of $O(EN + N^2 \log N)$ [24] and when $M \neq N$, with a complexity of $O(EK + K^2 \log(\min\{M, N\}))$ [22]. Once the matchings $\hat{\mathcal{H}}$ of (9) and (27) are identified, the matchings corresponding to (11) and (29) can also be identified in polynomial time. A naive algorithm for this would be to sequentially repeat the same algorithms on the graphs obtained by removing edges appearing in

$\widehat{\mathcal{H}}$ one at a time from the graph \mathcal{G} . The minimum weight matching obtained in all repetitions would correspond to the minimum weight matching on the original graph \mathcal{G} that is not identical to $\widehat{\mathcal{H}}$. In many practical applications this step is unnecessary as rejecting all hypotheses is not acceptable. In such cases one can use the estimates of (9) and (27). A practical application of such a solution and experimental evaluation of the method is reported in [1] where $M = N = K \approx 1500$ and for $M = N = K \approx 47000$ in [25].

The proposed solution can be extended in many directions. An important generalization is with respect to the requirement that all sequences have the same sample size. In practice, this is often not the case. For example in problem **(P1)**, it might be the case that each string y_i has a length $n_i = \alpha_i n$ with $\alpha_i \geq 1$. In such a case it might be possible to extend the result of Theorem III.3 by adapting the definitions of $D(\mathcal{H}_\ell)$ in (8) to

$$D(\mathcal{H}_\ell) = \sum_{(i,j) \in \mathcal{M}_\ell} \alpha_j D(\Gamma_{y_j} \parallel \mu_i)$$

and redefining the threshold $\widetilde{\lambda}$ in the statement of the theorem to $\widetilde{\lambda} = \lambda - \frac{|Z| \sum_{i=1}^N \log(n_i + 1)}{n}$. With this definition the optimality result of Theorem III.3 is expected to hold. Moreover the maximum likelihood interpretation of Lemma III.2 continues to hold for $M \geq N = K$. Alternatively, if all $n_i \geq n$ and all n_i are approximately equal, then the test proposed in Theorem III.3 can still be used, and the probability of error and probability of rejection are expected to be lower than those expected when $n_i = n$ for all i .

Similarly, the solution to problem **(P2)** can be generalized to the scenario in which the sequences have distinct lengths by following Gutman [4]. Let $n_i^x = \alpha_i^x n$ denote the length of sequence $x_i \in \mathcal{S}_1$ and $n_j^y = \alpha_j^y n$ denote the length of sequence $y_j \in \mathcal{S}_2$ with $\alpha_i^x \geq 1$ and $\alpha_j^y \geq 1$ for all i, j . Then the definition of $D(\mathcal{H}_\ell)$ in (26) can be changed to

$$\begin{aligned} D(\mathcal{H}_\ell) = & \sum_{(i,j) \in \mathcal{M}_\ell} \frac{n_i^x}{n} D\left(\Gamma_{x_i} \parallel \frac{n_i^x \Gamma_{x_i} + n_j^y \Gamma_{y_j}}{n_i^x + n_j^y}\right) \\ & + \frac{n_j^y}{n} D\left(\Gamma_{y_j} \parallel \frac{n_i^x \Gamma_{x_i} + n_j^y \Gamma_{y_j}}{n_i^x + n_j^y}\right) \end{aligned}$$

and the threshold $\widetilde{\lambda}$ in the statement of Theorem III.3 can be changed to

$$\widetilde{\lambda} = \lambda - \frac{|Z| \left(\sum_{i=1}^M \log(n_i^x + 1) + \sum_{j=1}^N \log(n_j^y + 1) \right)}{n}.$$

With this definition the optimality result of Theorem IV.3 is expected to hold. It can be noted that

$$\frac{n_i^x \Gamma_{x_i} + n_j^y \Gamma_{y_j}}{n_i^x + n_j^y} = \Gamma_{t_{ij}} \text{ where } t_{ij} \text{ is the concatenation of } x_i \text{ and } y_j.$$

Throughout this paper we focused on source probability distributions supported on a finite alphabet \mathcal{Z} . It might be possible to extend some of these results to probability distributions on continuous alphabets. For the problem with known sources studied in Section III, we know from Lemma III.2, that the choices \hat{H} and \tilde{H} correspond to the maximum-likelihood hypothesis, and the second most likely hypothesis. Hence, even for continuous alphabets, these hypotheses can be identified using standard techniques [2]. However, we recall that the optimal test of Theorem III.3 requires us to compare $D(\tilde{H})$ to a threshold. For continuous distributions the empirical distributions are in general not absolutely continuous with respect to the true distributions and thus D is always ∞ . Thus the definition of the decision regions for the optimal test have to be modified by replacing the Kullback Leibler divergence with some appropriately defined function of the log-likelihood function and by setting the thresholds intelligently. A potential approach is to adapt the method proposed for binary hypothesis testing in [26] to multiple hypothesis. The analysis of the error exponents and rejection exponents in such continuous alphabet problems are typically performed using Cramer's theorem rather than Sanov's theorem. The error exponent result of (24) is expected to continue to hold for a test that always decides in favor of \hat{H} without rejection.

The results of Section IV are more difficult to generalize to source distributions on continuous alphabets, because, in general, the empirical distributions of all sequences are expected to have mutually disjoint supports. However, if the source distributions are constrained to lie in some parametric family, for example, an exponential family [2] such as the class of Gaussian distributions of unknown means and variances equal to unity, it might be possible to identify optimal procedures via the maximum generalized likelihood interpretation of Lemma IV.2. This idea of restricting to finite dimensional parametric families is similar to the dimensionality reduction approach prescribed in [10] for universal hypothesis testing. These ideas are also useful in applications in which the alphabet size $|\mathcal{Z}|$ is large. As described in [10], test statistics for hypothesis testing problems on large alphabets suffer from large variance for moderate sequence lengths, and thus lead to poor error probability performances. Dimensionality reduction techniques like those proposed in [10] are an effective technique to address these concerns.

The solutions proposed in this paper for i.i.d. sources can also be easily extended to finite memory Markov sources on finite alphabets following the approach in [4]. Furthermore, it is possible to study the weak-convergence behavior of the test statistics of (9) and (27) following the method outlined in [27]. Using such an analysis it is possible to estimate the error probabilities for finite sample sizes.

ACKNOWLEDGMENTS

The author thanks the anonymous reviewers for several helpful suggestions and Farid Movahedi Naini for helpful discussions. This research was supported by ERC Advanced Investigators Grant: Sparse Sampling: Theory, Algorithms and Applications SPARSAM no 247006.

APPENDIX

For proving the various results we need some new notation and a few lemmas. For any sequence $s \in \mathcal{Z}^n$ we use T_s to denote the *type class* of s , i.e., the set of all sequences of length n with the same empirical distribution as s . The following lemmas are well known. For proofs see [14]. The first lemma below is just a restatement of Lemma II.1.

Lemma A.1. *For every $p \in \mathcal{P}(\mathcal{Z})$ and every $s \in \mathcal{Z}^n$,*

$$\frac{1}{(n+1)^{|\mathcal{Z}|}} 2^{-nD(\Gamma_s \| p)} \leq \mathbb{P}_p(T_s) \leq 2^{-nD(\Gamma_s \| p)}$$

where \mathbb{P}_p denotes the probability measure when all observations in s are drawn i.i.d. according to law p . □

Lemma A.2. *For any sequence $s \in \mathcal{Z}^n$ and any $\nu \in \mathcal{P}(\mathcal{Z})$ we have*

$$\nu(s) \leq 2^{-nH(\Gamma_s)}.$$

□

The following lemma is easy to see.

Lemma A.3. *For finite set \mathcal{Z} , we have*

$$\sum_{s \in \mathcal{Z}^n} 2^{-n(H(\Gamma_s))} \leq (n+1)^{|\mathcal{Z}|}.$$

Proof: Let \mathcal{P}_n denote the set of all types with denominator n . Let $T(P)$ be the set of sequences in \mathcal{Z}^n with type P . We have

$$\sum_{s \in \mathcal{Z}^n} 2^{-n(H(\Gamma_s))} = \sum_{P \in \mathcal{P}_n} |T(P)| 2^{-nH(P)} \tag{38}$$

$$\stackrel{(a)}{\leq} \sum_{P \in \mathcal{P}_n} 1 = |\mathcal{P}_n| \stackrel{(b)}{\leq} (n+1)^{|\mathcal{Z}|} \tag{39}$$

where (a) and (b) follow from [14, Ch. 11]. ■

The following lemma is also required for some proofs.

Lemma A.4. For $\mu_1, \mu_2, \dots, \mu_N \in \mathcal{P}(\mathcal{Z})$ let $\mu^{\text{prod}} := \mu_1 \times \mu_2 \times \dots \mu_N$ denote the product distribution. For $\pi \in \mathcal{P}(\mathcal{Z}^N)$, let $\check{\pi}$ denote the product distribution obtained from the marginals of π , i.e., $\check{\pi} = \pi_{1,\cdot} \times \pi_{2,\cdot} \times \dots \pi_{N,\cdot}$, where $\pi_{k,\cdot}$ denote the marginal distribution of π with respect to the k -th component. Then we have

$$D(\check{\pi} \parallel \mu^{\text{prod}}) \leq D(\pi \parallel \mu^{\text{prod}}).$$

Proof: We know that

$$D(\pi \parallel \mu^{\text{prod}}) = E_{X_1, X_2, \dots, X_N} \frac{\log \pi(X_1, X_2, \dots, X_N)}{\prod_{i \in [N]} \mu_i(X_i)}$$

where (X_1, X_2, \dots, X_N) has joint distribution π . Simplifying we have

$$\begin{aligned} D(\pi \parallel \mu^{\text{prod}}) &= E_{X_1, X_2, \dots, X_N} \log \left(\frac{\pi(X_1, X_2, \dots, X_N)}{\prod_{i \in [N]} \pi_{i,\cdot}(X_i)} \right) \\ &\quad + \log \left(\frac{\prod_{i \in [N]} \pi_{i,\cdot}(X_i)}{\prod_{i \in [N]} \mu_i(X_i)} \right) \\ &= E_{X_1, X_2, \dots, X_N} \log \left(\frac{\pi(X_1, X_2, \dots, X_N)}{\prod_{i \in [N]} \pi_{i,\cdot}(X_i)} \right) \\ &\quad + \sum_{i \in [N]} D(\pi_{i,\cdot} \parallel \mu_i) \\ &= \sum_{i \in [N]} H(X_i) - H(X_1, X_2, \dots, X_N) + \sum_{i \in [N]} D(\pi_{i,\cdot} \parallel \mu_i) \\ &\geq \sum_{i \in [N]} D(\pi_{i,\cdot} \parallel \mu_i) = D(\check{\pi} \parallel \mu^{\text{prod}}). \end{aligned}$$

where $H(\cdot)$ denotes Shannon entropy and the inequality follows a well known information theoretic inequality between the joint Shannon entropy of random variable and the sum of their individual Shannon entropies [14]. ■

A. Proof of Lemma II.3

Let $A = \{(x, y) : x \in \mathcal{Y}^n, y \in \mathcal{Y}^n, \text{ and } D(\Gamma_x \| \frac{1}{2}(\Gamma_x + \Gamma_y)) + D(\Gamma_y \| \frac{1}{2}(\Gamma_x + \Gamma_y)) > \lambda\}$. Then we have

$$\begin{aligned}
& \mathbb{P}\{(y_1, y_2) \in A\} \\
&= \sum_{(x, y) \in A} \nu(x) \nu(y) \\
&\stackrel{(a)}{\leq} \sum_{(x, y) \in A} 2^{-2nH(\frac{1}{2}(\Gamma_x + \Gamma_y))} \\
&= \sum_{(x, y) \in A} 2^{-n(H(\Gamma_x) + H(\Gamma_y))} \\
&\quad 2^{-n(D(\Gamma_x \| \frac{1}{2}(\Gamma_x + \Gamma_y)) + D(\Gamma_y \| \frac{1}{2}(\Gamma_x + \Gamma_y)))} \\
&\stackrel{(b)}{\leq} \sum_{(x, y) \in A} 2^{-n(H(\Gamma_x) + H(\Gamma_y) + \lambda)} \\
&\leq 2^{-n\lambda} \sum_{x \in \mathcal{Y}} 2^{-nH(\Gamma_x)} \sum_{y \in \mathcal{Y}} 2^{-nH(\Gamma_y)} \\
&\stackrel{(c)}{\leq} 2^{-n\lambda} (n+1)^{2|\mathcal{Y}|}
\end{aligned}$$

where (a) follows from Lemma A.2 applied to a concatenation of x and y , (b) from the definition of A , and (c) from Lemma A.3. Thus

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}\{(y_1, y_2) \in A\} \geq \lambda.$$

B. Proof of Lemma III.1

Consider an arbitrary tuple of sequences (y_1, y_2, \dots, y_N) . Let $T = (T_{y_1}, \dots, T_{y_N})$ denote the joint type-class of all the sequences, i.e., it is the set of all tuples of sequences with the same joint type as (y_1, y_2, \dots, y_N) :

$$T = \{(z_1, \dots, z_N) : z_i \subset \mathcal{Z}^n \text{ and } \Gamma_{z_i} = \Gamma_{y_i} \text{ for all } i \in [N]\}.$$

Any $(y'_1, y'_2, \dots, y'_N) \in T$ belongs to exactly one of the sets $\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R$. We modify the decision rule Ω as follows. For any joint type T we let Λ_ℓ include T if Ω_ℓ contains the most number of the sequences of T , for $\ell \in \{1, 2, \dots, J, R\}$. In case of ties we break them arbitrarily and include T in exactly one of the Λ_ℓ 's.

Let q_i^y denote the probability distribution of the source that produced sequence y_i under hypothesis \mathcal{H}_ℓ . For any hypothesis \mathcal{H}_ℓ with $\ell \in [J]$ and any joint type $T \subset \Lambda_k$ with $k \in [J] \cup \{R\}$ we have by

Lemma A.1 and definition of Λ_ℓ :

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_\ell}\{\Omega_k\} &\geq \mathbb{P}_{\mathcal{H}_\ell}\{\Omega_k \cap T\} \geq \frac{1}{J+1} \mathbb{P}_{\mathcal{H}_\ell}\{T\} \\
&\stackrel{(a)}{\geq} \frac{2^{-n(\delta(n) + \sum_{j=1}^N D(\Gamma_{y_j} \| q_j^y))}}{J+1}
\end{aligned} \tag{40}$$

where (a) follows via the first inequality in the statement of Lemma A.1 with $\delta(n) = \frac{N|Z|\log(n+1)}{n}$.

Combining the above result along with the definition of Λ_ℓ and Lemma A.1, we have

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_\ell}\{\Lambda_k\} &= \sum_{T: T \subset \Lambda_k} \mathbb{P}_{\mathcal{H}_\ell}\{T\} \\
&\stackrel{(a)}{\leq} \sum_{T: T \subset \Lambda_k} 2^{-n(\sum_{j=1}^N D(\Gamma_{y_j} \| q_j^y))} \\
&\stackrel{(b)}{\leq} \sum_{T: T \subset \Lambda_k} 2^{n\delta(n)} (J+1) \mathbb{P}_{\mathcal{H}_\ell}\{\Omega_k\} \\
&\leq \tau_n 2^{n\delta(n)} (J+1) \mathbb{P}_{\mathcal{H}_\ell}\{\Omega_k\}
\end{aligned}$$

where (a) follows via the second inequality in Lemma A.1 and (b) via (40). The quantity τ_n represents the number of joint types of length n . Since $\frac{\log \tau_n}{n} \rightarrow 0$ [14] and $\delta(n) \rightarrow 0$ we obtain the inequality relations claimed in the lemma by choosing $k \in [J]$ and $k = R$.

C. Proof of Lemma III.2

Under \mathcal{H}_ℓ let

$$I_y^\ell = \{j : \text{No edge in } M_\ell \text{ is incident on } y_j\}.$$

Also let q_i^y denote the probability distribution of the source that produced sequence y_i . We have

$$\begin{aligned}
& \arg \max_{\ell \in [J]} \max_{\substack{\mathcal{N} \subset \mathcal{P}(\mathcal{Z}) \\ \mathcal{M} \cap \mathcal{N} = \mathcal{K}}} \mathbf{P}_{\mathcal{H}_\ell}(y_1, y_2, \dots, y_N) \\
&= \arg \max_{\ell \in [J]} \sum_{(i,j) \in \mathbf{M}_\ell} \log \prod_{k=1}^n \mu_i(y_j(k)) \\
&\quad + \sum_{j \in I_y^\ell} \max_{q_j^y \in \mathcal{P}(\mathcal{Z})} \log \prod_{k=1}^n q_j^y(y_j(k)) \\
&\stackrel{(a)}{=} \arg \max_{\ell \in [J]} \sum_{(i,j) \in \mathbf{M}_\ell} \left(\sum_{z \in \mathcal{Z}} n \Gamma_{y_j}(z) \log(\mu_i(z)) \right) \\
&\quad + \sum_{j \in I_y^\ell} \log \prod_{k=1}^n \Gamma_{y_j}(y_j(k)) \\
&= \arg \max_{\ell \in [J]} \sum_{(i,j) \in \mathbf{M}_\ell} (-H(\Gamma_{y_j}) - D(\Gamma_{y_j} \parallel \mu_i)) \\
&\quad - \sum_{j \in I_y^\ell} H(\Gamma_{y_j}) \\
&= \arg \min_{\ell \in [J]} \sum_{(i,j) \in \mathbf{M}_\ell} D(\Gamma_{y_j} \parallel \mu_i) + \sum_{j \in [N]} H(\Gamma_{y_j}) \\
&= \arg \min_{\ell \in [J]} D(\mathcal{H}_\ell)
\end{aligned}$$

where (a) follows from the fact that the likelihood of a string is maximized by the empirical distribution.

D. Proof of Theorem III.3

As before, let

$$I_y^\ell = \{j : \text{No edge in } \mathbf{M}_\ell \text{ is incident on } y_j\}$$

denote the indices of the $N - K$ sequences in \mathcal{S} that are produced by sources in $\mathcal{N} \setminus \mathcal{K}$. Similarly, let

$$I_\mu^\ell = \{j : \text{No edge in } \mathbf{M}_\ell \text{ is incident on } \mu_j\}.$$

denote the indices of the $M - K$ sources in $\mathcal{M} \setminus \mathcal{K}$. We continue to use q_i^y to denote the probability distribution of the source that produced sequence y_i . Let

$$\tilde{\Lambda}_\ell = \{\underline{y} : D(\mathcal{H}_\ell) \geq \tilde{\lambda}\}, \ell \in [J].$$

The probability of error of decision rule Λ under hypothesis \mathcal{H}_ℓ is given by

$$P_\Lambda(\text{err}/\mathcal{H}_\ell) = \mathbf{P}_{\mathcal{H}_\ell} \left\{ \underline{y} \in \bigcup_{\substack{k=1 \\ k \neq \ell}}^J \Lambda_k \right\} \leq \mathbf{P}_{\mathcal{H}_\ell}(\tilde{\Lambda}_\ell).$$

We observe that $D(\mathcal{H}_\ell)$ in the definition of $\tilde{\Lambda}_\ell$ is a sum of the Kullback Leibler divergences between the empirical distributions of each y_i and some μ_j . The empirical distributions of each y_i can be interpreted as the marginal of a joint empirical distribution of \underline{y} interpreted as a sequence of length n drawn from Z^N . We also note that $\tilde{\lambda} \rightarrow \lambda$ and $n \rightarrow \infty$. The result of (14) follows directly by applying Sanov's theorem [16] combined with the conclusion of Lemma A.4.

For proving (15) we observe that for any test based on empirical distributions, we have

$$2^{-\lambda n} \geq P_\Omega(\text{err}/\mathcal{H}_\ell) = \sum_{\cup_{k \neq \ell} \Omega_k} \prod_{j=1}^N q_j^y(y_j)$$

where we use $q_j^y(s)$ to denote the probability that sequence s was generated i.i.d. under law q_j^y . Simplifying further we have,

$$\begin{aligned} 2^{-\lambda n} &\geq \sum_{\cup_{k \neq \ell} \Omega_k} \prod_{i \in I_y^\ell} q_i^y(y_i) \prod_{j \notin I_y^\ell} q_j^y(y_j) \\ &\stackrel{(a)}{\geq} \sum_{T \subset \cup_{k \neq \ell} \Omega_k} 2^{-n \sum_{i \in I_y^\ell} (D(\Gamma_{y_i} \| q_i^y) + \delta(n))} \\ &\quad 2^{-n \sum_{j \notin I_y^\ell} (D(\Gamma_{y_j} \| q_j^y) + \delta(n))} \\ &\geq 2^{-n \sum_{i \in I_y^\ell} (D(\Gamma_{y'_i} \| q_i^y) + \delta(n))} \\ &\quad 2^{-n \sum_{j \notin I_y^\ell} (D(\Gamma_{y'_j} \| q_j^y) + \delta(n))} \end{aligned}$$

where (a) follows from Lemma A.1 with $T = (T_{y_1}, \dots, T_{y_N})$ and $\delta(n) = \frac{|Z| \log(n+1)}{n}$, and $(y'_1, y'_2, \dots, y'_N) \in \cup_{k \neq \ell} \Omega_k$, and all distributions in $\mathcal{N} \setminus \mathcal{K} \subset \mathcal{P}(Z)$ are arbitrary. If we specifically choose $\mathcal{N} \setminus \mathcal{K}$ such that $q_j^y = \Gamma_{y'_j}$ for all $j \in I_y^\ell$ we get

$$\lambda \leq \sum_{j \notin I_y^\ell} (D(\Gamma_{y'_j} \| q_j^y)) + N\delta(n)$$

which further implies that

$$\cup_{j \neq \ell} \Omega_j \subset \tilde{\Lambda}_\ell. \tag{41}$$

Now let

$$\hat{\Lambda}_\ell := \cap_{j \neq \ell} \tilde{\Lambda}_j.$$

Hence,

$$\cup_\ell \Lambda_\ell = \{\underline{y} : D(\tilde{\mathcal{H}}) \geq \tilde{\lambda}\} = \cup_\ell \hat{\Lambda}_\ell.$$

Combining with (41) we get

$$\hat{\Lambda}_\ell = \cap_{j \neq \ell} \tilde{\Lambda}_j \supset \cap_{j \neq \ell} \cup_{k \neq j} \Omega_k \supset \Omega_\ell$$

and thus

$$\Lambda_R^c = \cup_\ell \Lambda_\ell = \cup_\ell \hat{\Lambda}_\ell \supset \cup_\ell \Omega_\ell = \Omega_R^c.$$

Hence

$$\Lambda_R \subset \Omega_R.$$

E. Proof of Lemma IV.1

This proof is very similar to that of Lemma III.1. Let $(x_1, x_2, \dots, x_M, y_1, y_2, \dots, y_N)$ be an arbitrary tuple of sequences. Let $T = (T_{x_1}, \dots, T_{x_M}, T_{y_1}, \dots, T_{y_N})$ denote the joint type-class of all the sequences, defined similarly to the definition in the proof of Lemma III.1 as

$$\begin{aligned} T &= \{(w_1, \dots, w_M, z_1, \dots, z_N) : w_i, z_j \in Z^n, \Gamma_{w_i} = \Gamma_{x_i} \\ &\text{and } \Gamma_{z_j} = \Gamma_{y_j} \text{ for all } i \in [M], j \in [N]\}. \end{aligned}$$

Any $(x'_1, x'_2, \dots, x'_M, y'_1, y'_2, \dots, y'_N) \in T$ belongs to exactly one of the sets $\Omega_1, \Omega_2, \dots, \Omega_J, \Omega_R$. We modify the decision rule Ω as follows. For any joint type T we let Λ_ℓ include T if Ω_ℓ contains the most number of the sequences of T , for $\ell \in \{1, 2, \dots, J, R\}$. In case of ties we break them arbitrarily and include T in exactly one of the Λ_ℓ 's.

Under hypothesis \mathcal{H}_ℓ , let q_i^x denote the probability distribution of the source that produced sequence x_i and q_i^y denote the probability distribution of the source that produced sequence y_i . For any hypothesis $\ell \in [J]$ and any joint type $T \subset \Lambda_k$ with $k \in [J] \cup \{R\}$ we have by Lemma A.1 and definition of Λ_ℓ :

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_\ell}\{\Omega_k\} &\geq \mathbb{P}_{\mathcal{H}_\ell}\{\Omega_k \cap T\} \geq \frac{1}{J+1} \mathbb{P}_{\mathcal{H}_\ell}\{T\} \\ &\geq \frac{2^{-n(\delta(n) + \sum_{i=1}^M D(\Gamma_{x_i} \| q_i^x) + \sum_{j=1}^N D(\Gamma_{y_j} \| q_j^y))}}{J+1} \end{aligned}$$

where $\delta(n) = \frac{(M+N)|Z|\log(n+1)}{n}$. Combining the above result along with the definition of Λ_k and Lemma A.1, we have

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_\ell}\{\Lambda_k\} &= \sum_{T \subset \Lambda_k} \mathbb{P}_{\mathcal{H}_\ell}\{T\} \\ &\leq \sum_{T \subset \Lambda_k} 2^{-n(\sum_{i=1}^M D(\Gamma_{x_i} \| q_i^x) + \sum_{j=1}^N D(\Gamma_{y_j} \| q_j^y))} \\ &\leq \sum_{T \subset \Lambda_k} 2^{n\delta(n)} (J+1) \mathbb{P}_{\mathcal{H}_\ell}\{\Omega_k\} \\ &\leq \tau_n 2^{n\delta(n)} (J+1) \mathbb{P}_{\mathcal{H}_\ell}\{\Omega_k\} \end{aligned}$$

where τ_n represents the number of joint types of length n . Since $\frac{\log \tau_n}{n} \rightarrow 0$ [14] and $\delta(n) \rightarrow 0$ the results follow by choosing $k \in [J]$ and $k = R$.

F. Proof of Lemma IV.2

We know that there are $M - K$ sequences in \mathcal{S}_1 and $N - K$ sequences in \mathcal{S}_2 that are not produced by sources in \mathcal{K} . We represent the indices of these sequences under hypothesis \mathcal{H}_ℓ by the following notation

$$I_x^\ell = \{j : \text{No edge in } \mathbf{M}_\ell \text{ is incident on } x_j\} \quad (42)$$

and

$$I_y^\ell = \{j : \text{No edge in } \mathbf{M}_\ell \text{ is incident on } y_j\}. \quad (43)$$

Furthermore, we let q_i^x denote the probability distribution of the source that produced sequence x_i and q_i^y denote the probability distribution of the source that produced sequence y_i .

We first observe that if $x, y \in \mathcal{Z}^n$ are two length n strings drawn under the same distribution from $\mathcal{P}(\mathcal{Z})$, then the maximum likelihood distribution that produced it is given by $\frac{1}{2}(\Gamma_x + \Gamma_y)$, the empirical distribution of the concatenated string. In other words

$$\arg \max_{\mu \in \mathcal{P}(\mathcal{Z})} \mu(x)\mu(y) = \frac{1}{2}(\Gamma_x + \Gamma_y) \quad (44)$$

Now we have

$$\begin{aligned} & \log \mathbf{P}_{\mathcal{H}_\ell}(x_1, x_2, \dots, x_M, y_1, y_2, \dots, y_N) \\ &= \sum_{(i,j) \in \mathbf{M}_\ell} \sum_{z \in \mathcal{Z}} \log (q_i^x(z))^{n(\Gamma_{x_i}(z) + \Gamma_{y_j}(z))} \\ & \quad + \sum_{i \in I_x^\ell} \sum_{z \in \mathcal{Z}} \log (q_i^x(z))^{n\Gamma_{x_i}(z)} \\ & \quad + \sum_{j \in I_y^\ell} \sum_{z \in \mathcal{Z}} \log (q_j^y(z))^{n\Gamma_{y_j}(z)} \end{aligned}$$

By (44)

$$\begin{aligned} & \max_{\substack{\mathcal{M}, \mathcal{N} \subset \mathcal{P}(\mathcal{Z}) \\ |\mathcal{M} \cap \mathcal{N}| = K}} \log \mathbf{P}_{\mathcal{H}_\ell}(x_1, x_2, \dots, x_M, y_1, y_2, \dots, y_N) \\ &= \sum_{(i,j) \in \mathbf{M}_\ell} \sum_{z \in \mathcal{Z}} \log \left(\frac{\Gamma_{x_i}(z) + \Gamma_{y_j}(z)}{2} \right)^{n(\Gamma_{x_i}(z) + \Gamma_{y_j}(z))} \\ & \quad + \sum_{i \in I_x^\ell} \sum_{z \in \mathcal{Z}} \log (\Gamma_{x_i}(z))^{n\Gamma_{x_i}(z)} \\ & \quad + \sum_{j \in I_y^\ell} \sum_{z \in \mathcal{Z}} \log (\Gamma_{y_j}(z))^{n\Gamma_{y_j}(z)} \end{aligned}$$

Hence

$$\begin{aligned}
& \arg \max_{\ell \in [J]} \max_{\substack{\mathcal{M}, \mathcal{N} \subset \mathcal{P}(\mathcal{Z}) \\ |\mathcal{M} \cap \mathcal{N}| = K}} \log \mathbb{P}_{\mathcal{H}_\ell}(x_1, \dots, x_M, y_1, \dots, y_N) \\
&= \arg \min_{\ell \in [J]} \sum_{(i,j) \in \mathbf{M}_\ell} \left\{ D(\Gamma_{x_i} \| \frac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})) \right. \\
&\quad \left. + D(\Gamma_{y_j} \| \frac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})) \right\} \\
&\quad - \sum_{i \in [M]} H(\Gamma_{x_i}) - \sum_{j \in [N]} H(\Gamma_{y_j}) \\
&= \arg \min_{\ell \in [J]} \sum_{(i,j) \in \mathbf{M}_\ell} \left\{ D(\Gamma_{x_i} \| \frac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})) \right. \\
&\quad \left. + D(\Gamma_{y_j} \| \frac{1}{2}(\Gamma_{x_i} + \Gamma_{y_j})) \right\}
\end{aligned}$$

where the last step follows from the fact that the sum of the entropy terms is equal for all hypotheses.

The conclusion of (30) follows directly, and that of (31) by a similar argument.

G. Proof of Theorem IV.3

We use the notation of I_x^ℓ and I_y^ℓ introduced in (42) and (43). Furthermore, as before let q_i^x (respectively q_i^y) denote the probability distribution of the source that produced sequence x_i (y_i).

This proof is very similar to that of Theorem III.3. Define

$$\tilde{\Lambda}_\ell = \{(\underline{x}, \underline{y}) : D(\mathcal{H}_\ell) \geq \tilde{\lambda}\}, \ell \in [J].$$

Clearly,

$$\Lambda_j \subset \tilde{\Lambda}_\ell \text{ for all } j \neq \ell$$

and hence

$$\cup_{j \neq \ell} \Lambda_j \subset \cup_{j \neq \ell} \left(\cap_{k \neq j} \tilde{\Lambda}_k \right) \subset \tilde{\Lambda}_\ell.$$

Therefore,

$$\begin{aligned}
P_\Lambda(\text{err}/\mathcal{H}_\ell) &= \sum_{\cup_{k \neq \ell} \Lambda_k} \prod_{i \in I_x^\ell} q_i^x(x_i) \prod_{j \in I_y^\ell} q_j^y(y_j) \\
&\quad \prod_{e \in \mathbf{M}_\ell} q_{e_1}^x(x_{e_1}) q_{e_2}^y(y_{e_2}) \\
&\stackrel{(a)}{\leq} \sum_{\tilde{\Lambda}_\ell} \prod_{i \in I_x^\ell} q_i^x(x_i) \prod_{j \in I_y^\ell} q_j^y(y_j) \\
&\quad \prod_{e \in \mathbf{M}_\ell} q_{e_1}^x(x_{e_1}) q_{e_1}^x(y_{e_2})
\end{aligned} \tag{45}$$

where (a) follows from the fact that under \mathcal{H}_ℓ for all $e \in \mathbf{M}_\ell$ we have $q_{e_1}^x = q_{e_2}^y$. If $\mathbf{M}_\ell = \{e^1, e^2, \dots, e^K\}$, let

$$\overline{\Lambda}_\ell := \{(x_{e_1^1}, x_{e_1^2}, \dots, x_{e_1^K}, y_{e_2^1}, y_{e_2^2}, \dots, y_{e_2^K}) : (\underline{x}, \underline{y}) \in \widetilde{\Lambda}_\ell\}.$$

From the definition of $\widetilde{\Lambda}_\ell$ it is evident that for a fixed matching \mathbf{M}_ℓ , the membership of $(\underline{x}, \underline{y})$ in $\widetilde{\Lambda}_\ell$ depends only on $\overline{\Lambda}_\ell$. Hence (45) becomes:

$$\begin{aligned} P_\Lambda(\text{err}/\mathcal{H}_\ell) &\leq \sum_{\overline{\Lambda}_\ell} \prod_{e \in \mathbf{M}_\ell} q_{e_1}^x(x_{e_1}) q_{e_1}^x(y_{e_2}) \\ &\stackrel{(b)}{\leq} \sum_{\overline{\Lambda}_\ell} \prod_{e \in \mathbf{M}_\ell} 2^{-2nH(\frac{1}{2}(\Gamma_{x_{e_1}} + \Gamma_{y_{e_2}}))} \\ &= \sum_{\overline{\Lambda}_\ell} 2^{-2n \sum_{e \in \mathbf{M}_\ell} H(\frac{1}{2}(\Gamma_{x_{e_1}} + \Gamma_{y_{e_2}}))} \end{aligned}$$

where (b) follows from Lemma A.2. Note that

$$\begin{aligned} 2H(\frac{1}{2}(\Gamma_{x_{e_1}} + \Gamma_{y_{e_2}})) &= H(\Gamma_{x_{e_1}}) + H(\Gamma_{y_{e_2}}) \\ &\quad + D(\Gamma_{x_{e_1}} \parallel \frac{1}{2}(\Gamma_{x_{e_1}} + \Gamma_{y_{e_2}})) \\ &\quad + D(\Gamma_{y_{e_2}} \parallel \frac{1}{2}(\Gamma_{x_{e_1}} + \Gamma_{y_{e_2}})) \end{aligned}$$

Thus

$$\begin{aligned} P_\Lambda(\text{err}/\mathcal{H}_\ell) &\leq \sum_{\overline{\Lambda}_\ell} 2^{-n\tilde{\lambda} - n \sum_{e \in \mathbf{M}_\ell} (H(\Gamma_{x_{e_1}}) + H(\Gamma_{y_{e_2}}))} \\ &\stackrel{(c)}{\leq} 2^{-n\tilde{\lambda}} \left((n+1)^{|\mathbf{Z}|} \right)^{2|\mathbf{M}_\ell|} \\ &= 2^{-n\tilde{\lambda}} (n+1)^{2K|\mathbf{Z}|} \\ &\leq 2^{-n(\lambda + O(\frac{\log n}{n}))} \end{aligned}$$

where (c) follows from Lemma A.3. This proves (33). This proof can be interpreted as an extension of Lemma II.3 to K pairs of empirical distributions.

For proving (34) we observe that for any test based on empirical distributions, we have

$$\begin{aligned}
2^{-\lambda n} &\geq P_{\Omega}(\text{err}/\mathcal{H}_{\ell}) \\
&= \sum_{\cup_{k \neq \ell} \Omega_k} \prod_{i=1}^M q_i^x(x_i) \prod_{j=1}^N q_j^y(y_j) \\
&= \sum_{\cup_{k \neq \ell} \Omega_k} \prod_{i \in I_x^{\ell}} q_i^x(x_i) \prod_{j \in I_y^{\ell}} q_j^y(y_j) \\
&\quad \prod_{e \in \mathbf{M}_{\ell}} q_{e_1}^x(x_{e_1}) q_{e_1}^x(y_{e_2}) \\
&\stackrel{(a)}{\geq} \sum_{T \subset \cup_{k \neq \ell} \Omega_k} 2^{-n \sum_{i \in I_x^{\ell}} (D(\Gamma_{x_i} \| q_i^x) + \delta(n))} \\
&\quad 2^{-n \sum_{j \in I_y^{\ell}} (D(\Gamma_{y_j} \| q_j^y) + \delta(n))} \\
&\quad 2^{-n \sum_{e \in \mathbf{M}_{\ell}} (D(\Gamma_{x_{e_1}} \| q_{e_1}^x) + D(\Gamma_{y_{e_2}} \| q_{e_1}^x) + 2\delta(n))} \\
&\geq 2^{-n \sum_{i \in I_x^{\ell}} (D(\Gamma_{x'_i} \| q_i^x) + \delta(n))} \\
&\quad 2^{-n \sum_{j \in I_y^{\ell}} (D(\Gamma_{y'_j} \| q_j^y) + \delta(n))} \\
&\quad 2^{-n \sum_{e \in \mathbf{M}_{\ell}} (D(\Gamma_{x'_{e_1}} \| q_{e_1}^x) + D(\Gamma_{y'_{e_2}} \| q_{e_1}^x) + 2\delta(n))}
\end{aligned}$$

where (a) follows from Lemma A.1 with $T = (T_{x_1}, \dots, T_{x_M}, T_{y_1}, \dots, T_{y_N})$ and $\delta(n) = \frac{|Z| \log(n+1)}{n}$, and $(x'_1, x'_2, \dots, x'_M, y'_1, y'_2, \dots, y'_N) \in \cup_{k \neq \ell} \Omega_k$ and all distributions in $\mathcal{U} \subset \mathcal{P}(Z)$ are arbitrary. If we specifically choose \mathcal{U} such that $q_{e_1}^x = \frac{1}{2}(\Gamma_{x'_{e_1}} + \Gamma_{y'_{e_2}})$ for all $e \in \mathbf{M}_{\ell}$, and $q_i^x = \Gamma_{x'_i}$ for all $i \in I_x^{\ell}$ and $q_j^y = \Gamma_{y'_j}$ for all $j \in I_y^{\ell}$ we get

$$\begin{aligned}
\lambda &\leq \sum_{e \in \mathbf{M}_{\ell}} \left(D(\Gamma_{x'_{e_1}} \| \frac{1}{2}(\Gamma_{x'_{e_1}} + \Gamma_{y'_{e_2}})) \right. \\
&\quad \left. + D(\Gamma_{y'_{e_2}} \| \frac{1}{2}(\Gamma_{x'_{e_1}} + \Gamma_{y'_{e_2}})) \right) + (M + N)\delta(n)
\end{aligned}$$

which further implies that

$$\cup_{j \neq \ell} \Omega_j \subset \tilde{\Lambda}_{\ell}. \quad (46)$$

Now let

$$\hat{\Lambda}_{\ell} := \cap_{j \neq \ell} \tilde{\Lambda}_j.$$

Hence,

$$\cup_{\ell} \Lambda_{\ell} = \{(\underline{x}, \underline{y}) : D(\tilde{\mathcal{H}}) \geq \tilde{\lambda}\} = \cup_{\ell} \hat{\Lambda}_{\ell}.$$

Combining with (46) we get

$$\hat{\Lambda}_\ell = \cap_{j \neq \ell} \tilde{\Lambda}_j \supset \cap_{j \neq \ell} \cup_{k \neq j} \Omega_k \supset \Omega_\ell$$

and thus

$$\Lambda_R^c = \cup_\ell \Lambda_\ell = \cup_\ell \hat{\Lambda}_\ell \supset \cup_\ell \Omega_\ell = \Omega_R^c.$$

Hence

$$\Lambda_R \subset \Omega_R.$$

REFERENCES

- [1] J. Unnikrishnan and F. M. Naini, “De-anonymizing private data by matching statistics,” in *Proc. of 51st Allerton Conference on Communication, Control, and Computing.*, Monticello, Illinois, 2013.
- [2] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, 3rd ed., ser. Springer Texts in Statistics. New York: Springer, 2005.
- [3] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. New York: Springer, 2009.
- [4] M. Gutman, “Asymptotically optimal classification for multiple tests with empirically observed statistics,” *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, 1989.
- [5] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in *Proc. 2008 IEEE Symposium on Security and Privacy*, Washington, DC, USA, 2008.
- [6] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the Crowd: The privacy bounds of human mobility,” *Scientific Reports*, vol. 3, Mar. 2013.
- [7] W. Hoeffding, “Asymptotically optimal tests for multinomial distributions,” *Ann. Math. Statist.*, vol. 36, pp. 369–408, 1965.
- [8] R. E. Blahut, “Hypothesis testing and information theory,” *IEEE Trans. Information Theory*, vol. IT-20, pp. 405–417, 1974.
- [9] G. Tusnády, “On asymptotically optimal tests,” *The Annals of Statistics*, vol. 5, no. 2, pp. pp. 385–393, 1977. [Online]. Available: <http://www.jstor.org/stable/2958993>
- [10] J. Unnikrishnan, D. Huang, S. Meyn, A. Surana, and V. Veeravalli, “Universal and composite hypothesis testing via mismatched divergence,” *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1587–1603, March 2011.
- [11] R. Ahlswede and I. Wegener, *Search Problems*. Chichester, U.K.: Wiley, 1987, German original by Teubner, 1979.
- [12] R. Ahlswede and E. Haroutunian, “On logarithmically asymptotically optimal testing of hypotheses and identification,” in *General Theory of Information Transfer and Combinatorics*, ser. Lecture Notes in Computer Science, R. Ahlswede, L. Bumer, N. Cai, H. Aydinian, V. Blinovskiy, C. Deppe, and H. Mashurian, Eds. Springer Berlin Heidelberg, 2006, vol. 4123, pp. 553–571.
- [13] D. B. West, *Introduction to Graph Theory*. Englewood Cliffs, N. J.: Prentice Hall, 2001, vol. 2.
- [14] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.
- [15] I. N. Sanov, “On the probability of large deviations of random magnitudes,” *Mat. Sb. N. S.*, vol. 42 (84), pp. 11–44, 1957.
- [16] A. Dembo and O. Zeitouni, *Large Deviations Techniques And Applications*, 2nd ed. New York: Springer-Verlag, 1998.

- [17] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*, 1st ed. Springer Publishing Company, Incorporated, 2008.
- [18] R. Blahut, "Hypothesis testing and information theory," *Information Theory, IEEE Transactions on*, vol. 20, no. 4, pp. 405–417, Jul 1974.
- [19] C. Leang and D. Johnson, "On the asymptotics of M-hypothesis Bayesian detection," *Information Theory, IEEE Transactions on*, vol. 43, no. 1, pp. 280–282, Jan 1997.
- [20] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. [Online]. Available: <http://dx.doi.org/10.1002/nav.3800020109>
- [21] L. Ramshaw and R. Tarjan, "A weight-scaling algorithm for min-cost imperfect matchings in bipartite graphs," in *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, Oct 2012, pp. 581–590.
- [22] —, "On minimum-cost assignments in unbalanced bipartite graphs," HP Labs, HP Labs technical report HPL-2012-40R1, 2012.
- [23] W. Cook, W. Cunningham, W. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*, ser. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2011.
- [24] M. L. Fredman and R. E. Tarjan, "Fibonacci heaps and their uses in improved network optimization algorithms," *J. ACM*, vol. 34, no. 3, pp. 596–615, Jul. 1987. [Online]. Available: <http://doi.acm.org/10.1145/28869.28874>
- [25] F. Movahedi Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Histograms as Fingerprints: User Identification by Matching Statistics," 2014. [Online]. Available: <http://infoscience.epfl.ch/record/201906>
- [26] T. S. Han, "Hypothesis testing with the general source," *Information Theory, IEEE Transactions on*, vol. 46, no. 7, pp. 2415–2427, Nov 2000.
- [27] J. Unnikrishnan and D. Huang, "Weak convergence analysis of asymptotically optimal hypothesis tests," 2013, submitted to *IEEE Transactions on Information Theory*.